1993

Bootstrap applications in proportional hazards models

Thomas Michael Loughin
Iowa State University

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9334996

Bootstrap applications in proportional hazards models

Loughin, Thomas Michael, Ph.D.

Iowa State University, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Bootstrap applications in proportional hazards models

by

Thomas Michael Loughin

A dissertation Submitted to the

Graduate Faculty in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Department: Statistics
Major: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa
1993

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

The research presented in this dissertation was partially supported by the National Institutes of Health through grant 5 R01 CA51831-02 from the National Cancer Institute. I wish to thank K. J. Koehler for his numerous helpful suggestions, for enlightening conversations, for his hours of review work, and for his general support, all of which have greatly improved the quality of this dissertation and the person who wrote it. Thanks are extended also to the faculty and staff of the Department of Statistics, and in particular to Professors P. N. Hinz and D. F. Cox, whose clear statistical philosophy has had an enormous impact on my own. Finally, and most importantly, an enormous thank-you goes to my wife, Marie, who, along with our daughter, Hannah, kept me well-nourished both physically and spiritually throughout the creation of this document.

GENERAL INTRODUCTION

Overview

A common problem in many fields of research is the analysis of the relationship of a variety of factors to *event times*. In medicine, there is often a need to assess the effects of various treatments on the time until an event occurs. Examples include time in remission for leukemia patients receiving two different drug therapies (Cox, 1972); survival times of patients accepted into a heart transplant program (Crowley and Hu, 1977); and time until viral positivity in blood serum samples collected for consecutive months in AIDS patients (Wei, Lin, and Weissfeld, 1989). In wildlife management studies, survival times of animals of a given species may be linked to factors such as sex or habitat type. In engineering, items in an experiment may be put under different levels of stress in an effort to study how those levels affect the lifetimes of the components. Books by Kalbfleisch and Prentice (1980), Lawless (1982), and Cox and Oakes (1984) contain many more examples from a variety of applications.

One complication that is common in many studies of event times is the presence of *censoring*. An observation of an event time is said to be censored if the exact time of the event is unknown, within the precision of measurement capability. Instead,

it may be known that the event occurred prior to a certain time (left censoring), within an interval of possible times (interval censoring), or after a certain time (right censoring). It is this last form of censoring that is most common in many studies of event times.

Several types of right censoring (hereinafter referred to simply as "censoring") are common in event time studies. Often data must be analyzed before all of the potential events have occurred. In studies of death times, for instance, it may be neither feasible nor ethical to withhold the results of the study until after all subjects have died. For this reason, individuals in the study are often subjected to censoring at a common fixed time. This kind of censoring is called *Type I* censoring.

Sometimes failure times are not observed because individuals leave the study before the final inspection design. These individuals are often said to be *lost to follow-up.* Factors such as treatment side effects, monitoring equipment failure, or individuals' moving to another location can lead to such uncontrollable censoring which is often placed under the general heading of *random censoring.*

An important assumption often made about random censoring is that its mechanism is unrelated to the mechanism of the event. In other words, it is assumed that the censoring of a subject provides no information about its remaining time until event occurrence. It is thus prohibited, for example, to remove subjects from the study because they appear to be near failure. Censoring adhering to this requirement is referred to as *independent censoring.*

Censored observations are not void of information about the event time of the subject. In particular, it is known that the subject's event *had not yet occurred* as of the last known follow-up time. A great deal of literature in event time analysis has

been devoted to the problems of handling censored data. See Lawless (1982, Section 1.4) and Kalbfleisch and Prentice (1980, Section 5.2) for detailed information about a variety of censoring mechanisms.

It is sometimes the case that times for more than one event are monitored for each observational unit in a study. This can happen in several ways:

- Events have the potential to recur, such as heart attacks or failures of reparable components;
- Each experimental/observational unit consists of multiple individuals, each of which is subject to a single (or multiple) event(s); matched-pair studies of twins and teratology studies of littermates are common examples;
- Different distinct events are monitored on the same individual, as with developmental milestones in young animals;
- Repeated measurements may be taken on an individual through subsampling some reproducible material (as in the blood serum viral positivity study analyzed in Wei, Lin, and Weissfeld (1989)).

Each different event may be subject to some form of censoring, which can complicate assessment of potential association between events. Kalbfleisch and Prentice (1980), Lawless (1982), and Cox and Oakes (1984) each contain a chapter on *multivariate* survival problems.

Organization of Dissertation

The first goal of the work presented here is to provide a method for improving the accuracy and precision of estimators under the proportional hazards model of Cox

(1972) for regression in survival analysis. This is accomplished by careful application of the bootstrap (Efron, 1979) to the case in which the explanatory variables are not random. Next, this methodology is extended to the problem of multivariate survival analysis. Specifically, using a working model for the estimation of regression parameters which assumes independence of the failure types a resampling plan for the underlying *joint* distribution is developed which provides appropriate variances and covariances for these estimates. Finally, a generalization of the first method is given which allows resampling from a *known* distribution, without full specification of the original distribution of the data.

To this end, the thesis is organized into primary sections as follows:

- The General Introduction, with Literature Review, appears first;
- Three papers, intended for submission to refereed publications, follow the introduction. These three papers address, in order, the three goals outlined above;
- A General Summary follows the papers, drawing together the results of the papers and suggesting directions for further research;
- References for the General Introduction, Literature Review, and General Summary appear after the General Summary. All references cited within the body of a paper are given at the end of that paper.

LITERATURE REVIEW

Univariate Survival

Let T be a continuous random variable corresponding to the time of the event under study, generically referred to here as "failure," and let $\boldsymbol{x}$ be a set of explanatory variables believed to influence the distribution of T. Call this conditional distribution $F(t;\boldsymbol{x})$, with its corresponding conditional *survivor function*,

$$S(t;\boldsymbol{x}) \equiv 1 - F(t;\boldsymbol{x}).$$

Define the conditional *hazard function* by

$$h(t;\boldsymbol{x}) \equiv \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t;\boldsymbol{x})}{\Delta t}, \tag{1}$$

where $h(t;\boldsymbol{x})\Delta t$ can be interpreted as the conditional probability of failure in $[t, t+\Delta t)$ given survival until time t. Note that for continuous variables T, $h(t;\boldsymbol{x})$ is strictly positive. Thus, we can define the conditional *cumulative hazard function*,

$$H(t;\boldsymbol{x}) \equiv \int_0^t h(u;\boldsymbol{x})du, \tag{2}$$

which is strictly increasing over the range of T.

Because of its heuristic appeal and its central role in determining failure time distributions, the hazard function has often been chosen as the focus of models for

relating explanatory variables to the failure time distribution. A wide variety of parametric and semiparametric models for $h(t;\boldsymbol{x})$ have been applied.

Weibull models

Among the more popular parametric models is the *Weibull distribution* (see, e.g., Lawless, 1982), whose hazard function has the form

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}, \tag{3}$$

where $\gamma > 0$ determines the shape of the distribution and $\lambda > 0$ determines its scale. A special case of the Weibull distribution is the *exponential distribution*, corresponding to $\gamma = 1$. Thus, the exponential distribution has constant hazard $h(t) = \lambda$.

To incorporate explanatory variables into the exponential distribution, Feigl and Zelen (1965) model the hazard as $\lambda = \lambda(\boldsymbol{x})$. More generally, this is done in (3) by writing

$$h(t;\boldsymbol{x}) = \lambda(\boldsymbol{x})\gamma(\lambda(\boldsymbol{x})t)^{\gamma-1}. \tag{4}$$

Note that when (4) holds, then the ratio of conditional hazards for individuals with explanatory variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is

$$\frac{h(t;\boldsymbol{x}_1)}{h(t;\boldsymbol{x}_2)} = \left(\frac{\lambda(\boldsymbol{x}_1)}{\lambda(\boldsymbol{x}_2)}\right)^{\gamma}, \tag{5}$$

which is constant across time. Lawless (1982) discusses the work of several authors which suggests that this mathematically convenient property holds reasonably well for many actual problems.

The proportional hazards model

Cox generalizes (5) in his landmark 1972 paper. He notes that the conditional hazard function can more generally be written as

$$h(t;\boldsymbol{x}) = h_0(t)g(\boldsymbol{x},\boldsymbol{\beta}), \tag{6}$$

where g is a strictly positive function of $\boldsymbol{x}$ and some unknown parameters $\boldsymbol{\beta}$, and $h_0(t)$ is the *baseline hazard function*, which is the hazard function for an individual with $g(\boldsymbol{x},\boldsymbol{\beta}) = 1$. The function g is known as the *relative risk function*, since it corresponds to the ratio of the hazard function for an individual with explanatory variables $\boldsymbol{x}$ to the baseline hazard function. A loglinear form,

$$g(\boldsymbol{x},\boldsymbol{\beta}) = e^{\boldsymbol{x}\boldsymbol{\beta}}$$

has been adopted by most authors.

Cox then proposes a likelihood function for estimating $\boldsymbol{\beta}$ *without specification of* h_O. His heuristic argument goes as follows. Let the data consist of n independent individuals. Order the failure times observed in the data, $t_{(1)},\ldots,t_{(k)}$, $k \leq n$, the possibility for $k < n$ resulting from right censoring. Label the corresponding explanatory variables $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k$. Define the *risk set*,

$$\mathcal{R}_i = \{j : t_j \geq t_{(i)}\},$$

as the set of all individuals known to be at risk of failure at time $t_{(i)}$.

Cox's likelihood is based on the premise that the probability that the failure at time $t_{(i)}$ happens to the individual with explanatory variables $\boldsymbol{x}_i$ is

$$Pr\{i \text{ fails at } t_{(i)} | \mathcal{R}_i, \text{someone fails at } t_{(i)}\}.$$

By the assumed continuity of T this becomes

$$\frac{\lim_{\Delta t \to 0} Pr\{i \text{ fails in } (t_{(i)}, t_{(i)} + \Delta) | i \text{ alive just prior to } t_{(i)}\}}{\sum_{l \in \mathcal{R}_i} \lim_{\Delta t \to 0} Pr\{l \text{ fails in } (t_{(i)}, t_{(i)} + \Delta) | l \text{ alive just prior to } t_{(i)}\}}$$

$$= \frac{h(t_i; \boldsymbol{x}_i)}{\sum_{l \in \mathcal{R}_i} h(t_i; \boldsymbol{x}_l)}$$

$$= \frac{g(\boldsymbol{x}_i, \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}_i} g(\boldsymbol{x}_l, \boldsymbol{\beta})} \tag{7}$$

by the factorization (6). Taking the product of (7) over the k failure times yields the likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{g(\boldsymbol{x}_i, \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}_i} g(\boldsymbol{x}_l, \boldsymbol{\beta})}. \tag{8}$$

Asymptotics in proportional hazards

It is suggested by Cox (1972) that this likelihood can be used in the same way as a proper likelihood. Estimates of $\boldsymbol{\beta}$ are obtained by maximizing (8), and the limiting normal distribution is used for inference on $\boldsymbol{\beta}$. The covariance matrix is estimated by the inverse of the observed Fisher information matrix, $I(\hat{\boldsymbol{\beta}}) = ((I_{rs}(\hat{\boldsymbol{\beta}})))$, where

$$I_{rs}(\hat{\boldsymbol{\beta}}) = -\frac{\partial^2 log L(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s}\Big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}}.$$

The use of L in (8) in this manner is questioned by several discussants of Cox (1972), since it is not a probability of any observable event. Kalbfleisch and Prentice (1973) justify (8) as a *marginal likelihood* (Kalbfleisch and Sprott, 1970) based on the ranks of the failure times when T is continuous. Cox (1975) also provides justification for (8) through his introduction of the *partial likelihood.*

The structure of such a likelihood requires that the data be expressible as

$(A_1, B_1, \ldots, A_n, B_n)$, allowing construction of a likelihood of the form,

$$\prod_{i=1}^{n} Pr(A_i|A^{(i-1)}, B^{(i-1)}; \theta) \times \prod_{i=1}^{n} Pr(B_i|A^{(i)}, B^{(i-1)}; \theta), \tag{9}$$

where $A^{(i)} = (A_1, \ldots, A_i)$, $B^{(i)} = (B_1, \ldots, B_i)$, and θ is a generic parameter label. The second of these products is what Cox calls the partial likelihood, denoted here by $\mathcal{L}(\theta)$. Note that it is neither a marginal nor a conditional likelihood, as defined in Kalbfleisch and Sprott (1970), so that theory developed for those forms does not necessarily hold.

Cox describes briefly the asymptotic behavior of estimators derived from this partial likelihood in the continuous case. The key element in his discussion is the score,

$$U = \frac{\partial log\mathcal{L}(\theta)}{\partial \theta} \equiv \sum_{i=1}^{n} U_i. \tag{10}$$

Assuming the usual regularity conditions, although here on the conditional density represented in terms of the second product of (9), he shows that the U_i's are uncorrelated, with mean 0 and variance equal to $I(\theta)$, minus the inverse of the second partial derivative of $log\mathcal{L}(\theta)$. Further regularity conditions on this conditional density allow the establishment of asymptotic normality for U and for $(\hat{\theta} - \theta)I^{1/2}(\hat{\theta})$.

The application of partial likelihood to the semiparametric estimation problem above is apparent upon letting A_i contain the censoring history in the interval $[t_{(i-1)}, t_{(i)})$ as well as the fact that a failure occurred at $t_{(i)}$, while B_i specifies the particular individual failing at $t_{(i)}$.

The problem of proving asymptotic results for estimators arising from the maximization of (8) has been studied by several authors. Tsiatis (1981), Andersen and Gill (1982), and Bailey (1983) are the most common citations in the literature. Ap-

plication of usual large sample results for maximum likelihood estimators, or more generally for M-estimators, is not possible here, due to the fact that the terms in (8) are not simply functions of *iid* variates. While their work focused on the single parameter case, as presented below, all authors gave extensions to the multiparameter situation.

Tsiatis begins by assuming that the explanatory variables are random with density $f_{\boldsymbol{X}}(\boldsymbol{x})$, and that the censoring times are both random (with a conditional hazard function possibly depending on $\boldsymbol{X}$) and conditionally independent of survival times, given the covariate $\boldsymbol{X}$. Further, it is required that these censoring times are bounded above by a T_0 such that $P(T > T_0) > 0$, where T represents the minimum of the survival and censoring times. Hence, the vectors $(T_i, \delta_i, \boldsymbol{X}_i)$ are considered as *iid* replicates, where δ_i is the censoring indicator for the i^{th} individual.

The proof of consistency of the one-dimensional parameter estimate $\hat{\beta}$ used by Tsiatis begins with lemmas and the strong law of large numbers to show that, for $\tilde{\beta}$ in a δ-neighborhood of β, the difference between the sample likelihood evaluated at β and that evaluated at $\tilde{\beta}$ converges almost surely to the corresponding difference in true likelihoods, as sample size tends toward ∞. This implies that the sample likelihood also has a maximum in the interior of the δ-neighborhood of β. Thus, by shrinking δ, the consistency of the maximizer of the sample likelihood is shown to be consistent for the true parameter β.

Proofs of asymptotic normality and the derivation of the variance of the limiting normal distribution as the inverse of minus the second derivative of the log likelihood follow from a Taylor expansion of the score function evaluated at $\hat{\beta}$. After writing this expansion as a sum of some associated processes, he shows the weak convergence of

$\sqrt{n}(\hat{\beta}-\beta)$ to a normal variate with the appropriate variance. The consistency of the variance estimator then results from the application of the aforementioned lemmas. Tsiatis also extends these results to the multiparameter case.

Bailey shows essentially the same results under a different set of conditions. He requires that both the explanatory variables and the censoring times be nonrandom. He also assumes that all of the explanatory variables are bounded by some constant and that the expected information matrix increases proportionally with n. The consistency of parameter estimates is then obtained by defining two sequences of events whose probabilities both approach zero together. In the first sequence the events consist of β and $\hat{\beta}$ differing by a distance of at least $n^{(-\frac{1}{2}+\epsilon)}$, while the second sequence has the likelihood evaluated at the same distance from β being greater than that *at* β. By showing that the elements of the first sequence are contained in the corresponding elements of the second, and that the probability of the second sequence goes to zero as n increases, the convergence of $\hat{\beta}$ to β in probability is proved. Proof of the consistency of the observed information matrix as an estimator of the true information matrix is again based on a Taylor expansion of the score function at $\hat{\beta}$.

The approach taken by Anderson and Gill is somewhat different in that they reformulate the problem as a counting process with intensity given by (6). They rely on the theory of multivariate counting processes, stochastic integrals, and local martingales to prove their results. Existence, uniqueness, and consistency of $\hat{\beta}$ and its variance estimator are shown in a manner similar to the approaches of the other authors. Normality results come from a special adaptation of Rebolledo's Central Limit Theorem for local square integrable martingales.

The problem of tied failure times

In practice, the assumption of a continuous distribution for the failure times is often violated. Tied failure times are often present in the data due to mild interval censoring or the finite precision of measurement. At present, it is unclear exactly what effect tied failure times have on the estimator $\hat{\boldsymbol{\beta}}$.

In fact, the likelihood $L(\boldsymbol{\beta})$ is not notationally adequate when ties are present, since the numerator in the product terms corresponds to the relative risk of a *single individual* failing at each ordered failure time. If many ties are present, then the problem might best be formulated as one occurring in discrete time. Cox (1972) and Kalbfleisch and Prentice (1973) are among those who have suggested models for this problem, they will not be discussed here. Note that it is only ties between *failures* that cause us difficulty. An observation which is censored at the same time at which a failure occurs is known to have lived at least that long, and so can logically be considered to follow the failure.

When ties are due to some slight grouping of what could otherwise be thought of as distinct times, the true likelihood can be found. However, it requires summing terms like those in (8), over all times at which ties occur and over all possible permutations of those times. With even small numbers of ties at a few failure times, this can be computationally cumbersome.

Operating with loglinear relative risk, Cox (1972), Peto (1972), Kalbfleisch and Prentice (1973), Breslow (1974), and Efron (1977) suggest methods for approximating the terms required in (8) in the presence of ties. All of these involve replacing $e^{\boldsymbol{x}_i\boldsymbol{\beta}}$ in the numerator with the product over the ties, $e^{\boldsymbol{s}_i\boldsymbol{\beta}}$, where $\boldsymbol{s}_i$ is the sum of the explanatory variables over the d_i individuals failing at $t_{(i)}$. The denomina-

tors are handled differently by different authors. All except Breslow take sums over smaller sets of permutations than would be required for the true likelihood. Breslow approximates these sums by

$$\left(\sum_{l \in \mathcal{R}_i} e^{\boldsymbol{x}_l \boldsymbol{\beta}} \right)^{d_i}.$$

This approximation has been adopted by most authors because of its computational ease.

There has been some concern regarding the adequacy of these approximations in the presence of heavily tied data. Farewell and Prentice (1980) focus on the problem of matched pairs in case-control studies. There, the small numbers of observations in a given stratum (as few as two) result in a large fraction of the stratum failing at each failure time. They find that the approximations due to Breslow and Efron do not perform well in this problem. Drawing the analogy to unstratified samples, they conclude that these approximations "should therefore not be used for case-control studies or for prospective studies in which the failure fractions at individual failure times are at all large" (p. 278).

All of the asymptotic results developed by the aforementioned authors for the proportional hazards model assume that the underlying distribution is continuous. Throughout this thesis as well, time is considered to be a continuous random variable. However, tied failure times are inevitably encountered in real life studies. Here, as in much of the proportional hazards literature, Breslow's (1974) approximation is used for handling ties. No attempt is made at present to determine the effect of tied failures on the methods put forth. This is a subject which requires further work.

Estimation of the survivor function

Once an estimate of β has been obtained, there may be interest in estimating the conditional survivor function, $S(t;\boldsymbol{x})$. By the proportional hazards factorization (6), it is easily seen that

$$S(t;\boldsymbol{x}) = [S_0(t)]^{g(\boldsymbol{x},\beta)}. \tag{11}$$

Hence, with the further estimation of $S_0(t)$, estimates of $S(t;\boldsymbol{x})$ can be obtained.

Cox (1972), Oakes (1972), Breslow (1972, 1974), and Kalbfleisch and Prentice (1973) all suggest estimators for $S_0(t)$ which resemble the product limit estimator of Kaplan and Meier (1958). The primary differences among them are how they specify the hazard function and deal with censoring between observed failure times.

Cox assumes, as an analog to Kaplan-Meier, that $h_0(t) = 0$ at all times between observed failures. He thus employs his discrete model to carry out maximum likelihood estimation at each distinct failure time. Maximization of this likelihood requires iterative estimation of parameters at each observed failure time.

In the discussion of Cox (1972), Oakes suggests letting the hazard be a slowly varying function of time, assuming it to be constant (but not zero) between failure times. His approach results in an estimate which can be evaluated analytically, and in which the locations of the censored observations between failure times are allowed to influence the estimate. In his reply, Cox notes that, "Mr. Oakes's suggestion appears superior to [my approach]" (p. 218).

Kalbfleisch and Prentice object to the fact that Cox's handling of the discrete case results in a likelihood designed to estimate a parameter from a logistic model instead of from the original continuous model. They develop a maximum likelihood estimator for the hazard function based on their own discrete model. Their result

requires iteration only at time points at which tied failure times occur.

These authors offer another estimator that provides a continuous (piecewise linear) survivor function. They create intervals independently of the observed failure times and assume constant hazard within each interval. This approach is quite similar to that of Oakes, except that Oakes fixes interval endpoints at the observed failure times.

Breslow's method is also similar to Oakes's method. He allows the hazard function to be constant and nonzero between failures and assumes all censoring takes place at the beginning of the interval. A maximum likelihood approach is developed that allows simultaneous estimation of β and the hazard. By integrating this estimated hazard function over time, he develops the baseline cumulative hazard estimator,

$$\hat{H}_0(t) = \sum_{i:t(i)<t} \left(\frac{d_i}{\sum_{l \in \mathcal{R}_i} e^{\boldsymbol{x}\hat{\boldsymbol{\beta}}}} \right). \tag{12}$$

From this, one can use either $\hat{S}_0(t) = exp(-H_0(t))$ or, upon noting that $e^{-u} \approx 1-u$ for small u,

$$\tilde{S}_0(t) = \prod_{i:t(i)<t} \left(\frac{1-d_i}{\sum_{l \in \mathcal{R}_i} e^{\boldsymbol{x}\hat{\boldsymbol{\beta}}}} \right).$$

All of the proposed methods can be expected to provide similar results in most cases (Kalbfleisch and Prentice, 1973). For all survivor function estimates in this thesis, Breslow's method is used.

Residuals in proportional hazards regression

It is well known that *residuals* can be used in linear models for assessing the adequacy of those models (see, for example, Neter, Wasserman, and Kutner, 1985).

Cox and Snell (1968) generalize the definition of residuals as follows.

Suppose there exists a random variable Y and a vector of parameters $\boldsymbol{\theta}$ such that

$$Y_i = q_i(\boldsymbol{\theta}, \epsilon_i), \ i = 1, \ldots, n, \tag{13}$$

where $\boldsymbol{\epsilon}$ is a vector of *iid* unobserved random variables. Suppose also that there exist functions v_i such that

$$\epsilon_i = v_i(\boldsymbol{\theta}, Y_i), \ i = 1, \ldots, n. \tag{14}$$

If $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$, then the generalized residual can be defined as

$$\hat{\epsilon_i} = v_i(\hat{\boldsymbol{\theta}}, Y_i), \ i = 1, \ldots, n. \tag{15}$$

Linear regression satisfies the requirements of this definition upon setting

$$q_i(\boldsymbol{\theta}, \epsilon_i) = \boldsymbol{X}_i\boldsymbol{\theta} + \epsilon_i$$

and

$$\hat{\epsilon_i} = Y_i - \boldsymbol{X}_i\hat{\boldsymbol{\theta}}.$$

The proportional hazards model (6) also has generalized residuals. Let

$$U = S(T; \boldsymbol{x}) = [S_0(T)]^{g(\boldsymbol{x}, \boldsymbol{\beta})}, \tag{16}$$

which corresponds to the probability integral transformation of T conditional on $\boldsymbol{x}$. This transformation is suggested in Section 10 of Cox and Snell. It is well known (e.g., DeGroot, 1975, p.127–128) that $U \sim$ Uniform(0,1). This U can be considered as a generalized residual with implicit model

$$T = S_0^{-1}(U^{1/g(\boldsymbol{x}, \boldsymbol{\beta})}).$$

Observed residuals can be found by evaluating (16) at observed failure times, using estimates of β and S_0.

Several authors have used estimates of the conditional cumulative hazard function as residuals in the proportional hazards model. Note that this corresponds to a minus-log transform of U and hence has a unit exponential distribution. Crowley and Hu (1977) plot cumulative hazard residuals based on Breslow's estimate (12) against a unit exponential curve to assess the adequacy of their proportional hazards assumption, and they plot these values against potential covariates to detect important but omitted variables. Kay (1977) calculates these residuals by the method of Kalbfleisch and Prentice (1973) for use in suggesting a parametric model for $h_0(t)$.

Residuals for censored data can be handled in several ways. The simplest is to treat them as censored observations from the residual distribution. In that case, an adjustment can be made to "uncensor" the residual. For instance, a censored exponential residual Y has as its true failure time distribution (conditional on survival until the censored value c) a unit exponential distribution,

$$Pr\{Y > y|Y > c\} = e^{-(y-c)}I(y > c). \tag{17}$$

Thus, an approximation to the true uncensored residual is obtained by adding to the residual its expected or median remaining life, 1 or $log2$, respectively. Both of these are done by Crowley and Hu (1977), who find the empirical properties of the latter adjustment more desirable. Similarly for uniform residuals (which are *left*-censored when T is right-censored),

$$Pr\{U > u|U > c\} = \frac{u}{c}I(u < c). \tag{18}$$

The residual can thus be adjusted to $\frac{1}{2}c$.

Another option for uncensoring residuals is to randomly select a value using the distributions (17) or (18). The appeal of this suggestion will become apparent when residual resampling plans are introduced.

Lagakos (1981) expresses concern that the residuals estimated by the above nonparametric methods do not possess the anticipated properties. He determines in particular, that rather than approximating a unit exponential variate, the cumulative hazard residual approximates the conditional expectation of a unit exponential given the ranks of the failures. He shows through enumeration for a sample of size three that the resulting expectations and variances may deviate substantially from those of a unit exponential, and hence recommends against using these residuals in a test of overall fit.

Crowley and Storer (1983) perform a small simulation study with a more realistic sample size of 20. They find that the means for the residuals are quite close to one, but the variances are just under 0.8. The average ordered residuals are generally slightly larger than those of a unit exponential.

As a follow-up to the work of Crowley and Storer a small Monte Carlo simulation has been run in conjunction with the present work. The goal is to examine the adequacy of Breslow's estimator for uniform residuals. The two-sample problem is considered, and samples of size 10, 20, 50, 100, and 1000 are used, covering the range of most likely sample sizes in real problems. Two different parameter values are used—1.5 and 0—corresponding to large and no treatment effects, respectively. For each parameter/sample size combination, 1000 data values are generated from a distribution with $h_0(t) = (1-t)^{-1}$, $0 < t < 1$ and $g(x,\beta) = e^{x\beta}$. (The selection of h_0 is chosen for convenience, since h_0 may be chosen arbitrarily, as will be seen

later.) All pseudo-random number generation is accomplished through the algorithm of Wichman and Hill (1982).

The results of this study indicate that the Breslow estimator tends to underestimate the expected uniform order statistics. However, any differences between the observed positions of the ordered residuals and their expected positions disappear quickly as sample size increases (the rate appears to be roughly n^{-1}). In samples as small as 20, the difference is only about .2 at the largest.

Perhaps more disturbing is an apparent large negative serial correlation between consecutive ordered residuals. This trend is particularly clear for smaller samples when $\beta = 1.5$. No explanation for this has been found in the literature. Further study of these residuals may be needed.

Small-sample studies

Despite its popularity in a wide variety of applications, the proportional hazards regression model has received little treatment in small-sample studies. The most common citation in the literature is to Johnson *et al.* (1982), who study the effects of sample size, type II censoring, and additional covariates on the bias and asymptotic variance of the maximum partial likelihood estimator $\hat{\beta}$. Another study by Costanza and Nichola (1982) investigates the effects of varying levels of random censoring on $\hat{\beta}$.

Costanza and Nichola use a single large data set, measuring age at death for 1395 people in Vermont who died from cardiovascular disease in 1974. A certain disease indicator, known to be highly significant, is used as their primary explanatory variable. They then impose random censoring on the sample at rates ranging from

5% to 90%. Based on their simulated censoring, they determine that the average location of the estimate $\hat{\beta}$ does not change with censoring, but that the variance of the 100 estimates increases by over 150 times in going from 5% to 90% censoring.

The study by Johnson *et al.* (1982) deals with smaller sample sizes more commonly encountered in experiments and clinical trials. These authors first summarize the results of their previous work, which describes the performance of the estimator $\hat{\beta}$ in the two-sample case. This earlier study considers the effects of sample size (20, 40, 60, 100), censoring (0, 50, or 80% type II), and unbalanced treatment assignment (samples split 50:50 or 20:80 between groups) on the bias and variance of the regression parameter estimator. They simulates data from an exponential distribution for their analyses.

The results of that study suggest that, for uncensored data with balanced assignment, biases in $\hat{\beta}$ range from around 7% at $n = 20$ to about 1.5% for $n = 100$. These figures are relatively consistent for true values of β ranging from .25 to 1. The asymptotic variance estimate typically underestimates the finite sample variance (the variance of the parameter estimates over the Monte Carlo trials) by 5% ($n = 100$) to 20% ($n = 20$), with greater underestimation for larger values of β.

When censoring is introduced, this underestimation of variance becomes more severe, though bias of $\hat{\beta}$ is not noticeably affected. Unbalanced treatment assignments, on the other hand, increase the biases of $\hat{\beta}$ by 30–60%, but the two variances stay "in reasonably close agreement" (p. 688).

Johnson *et al.* continue this work with an examination of the properties of estimates of a two-sample treatment effect in the presence of a covariate. They consider cases with continuous and discrete covariates, sample sizes of 40 and 100,

and possible type II censoring at 40%. Their results suggest that the relative bias of the treatment effect is roughly constant over varying values of β for each case. This relative bias changes only slightly with increasing values of the covariate coefficient, β_2. They also note an interaction between censoring and the value of β_2: when β_2 is low, censoring reduces bias, but when β_2 is high, an increase in bias is seen. No explanation is given for this effect. Finally, the asymptotic variance underestimates the finite sample variance by about 10% in most cases studied.

Monotone likelihood problems

When doing Monte Carlo studies of complex estimators, there is sometimes the possibility of computational difficulties in finding the estimate for certain data sets. Bryson and Johnson (1981) note that, in the proportional hazards setting, the parameter β is not identifiable if any of the explanatory variables (or certain functions thereof) is monotone with respect to the ordered *failure* times. In the two-sample problem, this is the case when the last failure observed in one group precedes the first failure in the other group. This results in a *monotone likelihood*; i.e., one that has no global maximum in $(-\infty, \infty)$.

Since the parameter estimates in monotone likelihood cases are $\pm\infty$ (i.e., iterative procedures for maximizing the likelihood do not converge), this causes potentially severe problems with computing and interpreting simulation results. Bryson and Johnson recommend detecting, omitting, and replacing such extreme samples *before estimation.* Then estimates of bias and finite sample variance are made conditional on the existence of finite parameter estimates. In the simulations and bootstraps in this thesis, all biases and variances are calculated in this manner.

Previous Work: Multivariate Survival Analysis

As described in the Overview, multivariate survival problems can take on a variety of structures. Here the problem of distinct events that can be modeled marginally will be emphasized, which includes any experiment in which the times to the occurrences of two or more distinct events are monitored on each subject, as well as some repeated measures experiments like the viral positivity example of Wei, Lin, and Weissfeld (1989). This is more general than the competing risks problem in that each subject may potentially experience all of the all of these events. The problem can be formulated as follows. Let

$$\boldsymbol{T} = (T_1, \ldots, T_M)$$

where T_m is the event time random variable for the m^{th} event type. Let there also exist a set of explanatory variables,

$$\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M),$$

where $\boldsymbol{x}_m$ is a $1 \times p_m$ vector. These may be the same for all margins, or they may differ arbitrarily. Define the *joint survivor function* by

$$S(\boldsymbol{t}; \boldsymbol{x}) \equiv Pr\{T_1 > t_1, \ldots, T_M > t_M; \boldsymbol{x}\}, \tag{19}$$

and the *marginal survivor functions* by

$$S_m(t_m; \boldsymbol{x}_m) \equiv Pr\{T_m > t_m; \boldsymbol{x}_m\} \; m = 1, \ldots, M.$$

(The term "conditional," for dependence on $\boldsymbol{x}$, is left out to avoid confusion.) Joint and marginal hazard, cumulative hazard, and distribution functions can be defined accordingly.

In a multivariate problem there may be several purposes to the analysis. The effects of the explanatory variables may be foremost; effects within margins or, in the case of repeated measures, changing effects across margins may be of interest. Estimation of marginal or joint survivor functions may be of importance. An assessment of the association among margins is often desired. Parametric and semiparametric methods have been developed for all of these problems. Hougaard (1987) provides an outstanding survey of these methods, with an emphasis on measuring dependencies among margins. In this thesis the focus is on estimating the effects of explanatory variables.

Parametric models

As with the univariate case, a number of parametric multivariate models have been proposed which can incorporate explanatory variables. Many of these are designed so that the margins are familiar univariate models. Distributions proposed by Gumbel (1960), Freund (1961), and Marshall and Olkin (1967) all have exponential margins, while those by Hougaard (1987) and Crowder (1989) simplify to Weibull marginals. One problem with many parametric multivariate models is the structure imposed on the association among margins. The distributions of Gumbel, Freund, Hougaard, and Marshall and Olkin all have constraints on the possible range of correlations, thus limiting their usefulness. Parameter estimation can be complicated by censoring, although it is usually straightforward, according to Lawless (1982, Section 10). Lawless also points out that some of these distributions have little motivation in terms of modeling realistic failure processes or providing a good representation for certain types of data. In fact such models are often presented without any indication

of how one might assess the adequacy of the model for a given data set. In Crowder's distribution, the null hypothesis of independence is not easily tested, since the corresponding parameter value renders another parameter unidentifiable.

Thus, it appears difficult to know which, if any, existing parametric models might best be used on a particular set of multivariate survival data. This is discouraging to those who wish to apply these methods.

Semiparametric models

Several authors have developed models which allow for the estimation of regression effects without complete specification of a distribution. Clayton and Cuzick (1985) develop a generalization of the distribution first introduced by Clayton (1978) and refined by Oakes (1982). The distribution is very general with respect to allowable marginal forms. Clayton and Cuzick introduce fixed explanatory variables into the distribution and suggest a likelihood by which corresponding regression parameters and the association parameter can be estimated.

Klein, Keiding, and Kamby (1989) generalize the distribution of Marshall and Olkin to allow for explanatory variables. They describe four models, all assuming a factorization similar to (6) for each hazard specified in the distribution. Likelihoods are developed for each model.

These semiparametric multivariate models provide an interesting alternative to full parametric specification. There is great potential for further development in this area.

Independence Working Model approach

When the primary concern in the analysis of multivariate failure time data is assessing the effects of explanatory variables on the times, models with complex association structures may not be necessary. Papers by Huster, Brookmeyer, and Self (1989) and Wei, Lin, and Weissfeld (1989) propose estimating regression parameters under a temporary assumption of independence of the margins. Huster *et al.* propose modeling paired data (disease status in two eyes) with independent Weibull distributions, while Wei *et al.* use marginal proportional hazards models. The likelihood resulting from such a formulation is called an *Independence Working Model* (IWM) by Huster *et al.*

The appeal in using an IWM is that the association, which is a nuisance to the estimation of the regression parameters, is completely eliminated from the estimation procedure. Goodness-of-fit and omitted-variable tests can be done within individual univariate margins, allowing the practitioner great flexibility in developing regression models. There is expected to be some loss of efficiency, however, since information contained in each margin cannot be used outside that margin. Huster *et al.* study the asymptotic relative efficiency of their IWM estimator when the times follow the bivariate distribution of Clayton (1978) and Oakes (1982). They find that the efficiency of their estimator is reasonably high when correlation is low, but that this efficiency decreases to below 50% when association becomes high. At a correlation of .64, the asymptotic relative efficiency ranges between 70–90% under varying conditions of practical interest.

As long as one can correctly specify the marginal distributions (which, in fully parametric models, is required anyway *in addition* to an association structure), the

parameter estimates obtained by maximizing the IWM likelihood are consistent (proofs are given in appendices in both papers), but inverting the information matrix from the IWM likelihood leads to inconsistent variance estimation in general. Royall (1986) applies a result developed by Huber (1967) to obtain a consistent estimator for the covariance matrix of the limiting normal distribution of the IWM parameter estimate that is "robust" to the incorrect specification of the likelihood.

Generically, let θ be a parameter, and let

$$L(\theta) \equiv \prod_{i=1}^{n} f_{\theta}(x_i)$$

be a likelihood function for estimating θ, $U_i(\theta)$ be the individual score

$$U_i(\theta) \equiv \frac{\partial}{\partial \theta} log f_{\theta}(x_i),$$

and $I(\theta)$ be the observed information matrix,

$$I(\theta) \equiv -\frac{\partial^2}{\partial \theta \partial \theta'} log L(\theta).$$

Then, under some regularity conditions on $f_{\theta}(x_i)$, Huber and Royall showed that the asymptotic variance of the estimator $\hat{\theta}$ maximizing $L(\theta)$ is consistently estimated by the "sandwich estimator"

$$V(\hat{\theta}) = I^{-1}(\hat{\theta}) \Lambda(\hat{\theta}) I^{-1}(\hat{\theta}), \tag{20}$$

where

$$\Lambda(\hat{\theta}) = \sum_{i=1}^{n} U_i(\hat{\theta}) [U_i(\hat{\theta})]'.$$

If the likelihood is correctly specified, $\Lambda(\hat{\theta}) = I^{-1}(\hat{\theta})$, so that (20) reduces to $I^{-1}(\hat{\theta})$. When the likelihood fails to adequately describe the data, the factor $\Lambda(\hat{\theta})$ provides an

adjustment which results in a consistent variance estimate. In related work, Lin and Wei (1989) use (20) in a univariate proportional hazards model as a model-robust estimator of variance.

All of the results reviewed thus far rely on large-sample theory for inference on the parameters. The development of inference procedures which do not rely on asymptotics is one of the main goals of this thesis. It is the multivariate survival problem which provides the primary motivation for the bootstrap methods which will be described later.

The Bootstrap

Large sample theory provides a foundation upon which most statistical inference is based. However, the adequacy of asymptotic results for describing the behavior of an estimator in a given sample has often been called into question. A review paper by Miller (1974) describes a technique called the *jackknife*, in which an estimator is calculated for all n subsamples of $n-1$ observations in the data set. Introduced in 1949 mainly for bias correction, this form of *resampling* provides appropriate bias-corrections and variance estimates in rather limited situations (Miller, 1974; Efron, 1982).

Another form of resampling, the *bootstrap*, has seen application in a much wider variety of problems. Efron (1979) developed the bootstrap as a means of estimating sampling distributions (and associated statistics) of estimators in small samples.

The idea behind the bootstrap begins with F_{θ}, the distribution function of interest in a given statistical problem, where θ is some parameter of interest. Let $G_n(\hat{\theta}|F_{\theta})$ be the sampling distribution of an estimator $\hat{\theta}$ based on simple random

samples of size n from F_{θ}. Let $\hat{F}_n$ be some approximation to the distribution F_{θ} (e.g., the empirical df or a parametric form evaluated at $\hat{\theta}$) based on the n observations. The bootstrap approximates $G_n(\hat{\theta}|F_{\theta})$ by $G_n(\theta^*|\hat{F}_n)$, where θ^* is an estimate of θ based on a simple random sample of size n from $\hat{F}_n$. While only one sample is available from the unknown F_{θ}, resulting in only one observation from $G_n(\hat{\theta}|F_{\theta})$, the fact that $\hat{F}_n$ is completely known theoretically allows any feature of $G_n(\theta^*|\hat{F}_n)$ to be evaluated. Typically, the form of $G_n(\theta^*|\hat{F}_n)$ is too complex for convenient analytical calculation, and approximations are obtained through Monte Carlo resampling of $\hat{F}_n$.

Theoretical considerations

The asymptotic validity of the bootstrap approximation has been shown for several estimation problems (Efron, 1979; Singh, 1981; Bickel and Freedman, 1981). Necessary to all of these is the idea that $G_n(\hat{\theta}|F_{\theta})$ must depend *smoothly* on F_{θ}; i.e., small changes in F_{θ} do not result in large changes in G_n. An example is given in Beran (1982) in which the bootstrap fails due to this lack of smoothness.

When the bootstrap is valid, it often provides a *better* estimate of $G_n(\hat{\theta}|F_{\theta})$ than does asymptotic theory. In particular, Hall (1992) emphasizes that when θ is a pivotal statistic and the asymptotic distribution of the original estimator is normal, the bootstrap provides an estimate of $G_n(\hat{\theta}|F_{\theta})$ whose error is of smaller order than is that of the limiting normal approximation.

Bootstrap uses

The bootstrap has found favor in problems where large-sample theory is either inadequate or unavailable. For such problems, questions regarding bias and variability

of the estimator can often be answered through computational rather than analytical power.

Bias correction was one of the original purposes of the jackknife (Miller, 1974) and is also a common task for the bootstrap. For simplicity, consider the case of a single parameter θ, and its estimator $\hat{\theta}$ based on a random sample from F_θ. We define bias by

$$R = E(\hat{\theta} - \theta | F_\theta), \tag{21}$$

where the expectation is taken under the true distribution of $\hat{\theta}$. This could hypothetically be found by sampling from F_θ infinitely often. Now define θ^* to be the same estimate of θ, but based on a simple random sample drawn from $\hat{F}_n$. Note that $\hat{\theta}$ can be considered as the *true* parameter value for the distribution $\hat{F}_n$. Then the bias of θ^* for estimating $\hat{\theta}$ is

$$R^* = E_*(\theta^* - \hat{\theta} | \hat{F}_n), \tag{22}$$

where now E_* is the expectation taken over $\hat{F}_n$. As before, this could be evaluated precisely through infinite sampling from $\hat{F}_n$ (although only a finite number of samples is needed if $\hat{F}_n$ is the empirical distribution). Computing θ^* for some sufficiently large number, B, of resamples provides us with

$$\bar{\theta}^* = \sum_{i=1}^{B} \theta_i^*,$$

an approximation to the expectation $E_*(\theta^*)$.

Assuming the bootstrap is valid, then R^* provides a reasonable approximation to R. Using our sole value of $\hat{\theta}$ as an approximation to its expectation, and combining (21) and (22) yields

$$\theta \approx \hat{\theta} - R^*.$$

Hence, a bias-corrected version of $\hat{\theta}$, say $\hat{\theta}_{BC}$, can be computed as

$$\hat{\theta}_{BC} = 2\hat{\theta} - \bar{\theta}^*.$$

Efron (1982) gives mathematical details of this procedure, but warns that it may actually increase the mean squared error due to increased variability in the resulting estimator. On the other hand, bias correction may result in a decrease in variability, thus reducing the mean squared error beyond the effect of the bias reduction. The utility of bias correction should be studied before it is employed.

The bootstrap estimates $\{\theta_i^*\}_{i=1}^n$ can also be used to compute an estimate of standard error. As with bias, this is done with the empirical estimate,

$$\sigma^* = \left(\frac{1}{B-1} \sum_{i=1}^{B} (\theta_i^* - \bar{\theta}^*)^2 \right)^{\frac{1}{2}}.$$

Confidence intervals

Much work has been done in the area of bootstrap confidence intervals. While it is sometimes possible to simply use the bootstrap estimates of bias and standard error in a normal-theory interval, this does not make full use of the information available from $G_n(\hat{\theta}|F_\theta)$.

Efron (1981b) introduces the *percentile method* of estimating confidence limits nonparametrically. Suppose a two-sided, $(1-\alpha) \times 100\%$ confidence interval is desired. Since $G_n(\theta^*|\hat{F}_n)$ represents a good nonparametric approximation for the true sampling distribution $G_n(\hat{\theta}|F_\theta)$, then the desired percentiles of $G_n(\hat{\theta}|F_\theta)$ can be estimated by the corresponding percentiles of $G_n(\theta^*|\hat{F}_n)$.

Notice that this procedure does not account for the potential bias inherent in the estimator $\hat{\theta}$. Efron (1981b) shows that this interval has the appropriate coverage

probability if there exists a monotone transformation $\phi = w(\theta)$ such that $\hat{\phi} - \phi$ and $\phi^* - \hat{\phi}$ have, for all θ, a common distribution which is symmetric about the origin. For cases where there is some bias effect, Efron suggests the *bias-corrected percentile method*, denoted by BC.

The BC method assumes that the transformation w suggested above leads to both $\hat{\phi} - \phi$ and $\phi^* - \hat{\phi}$ having, for all values of θ, the same normal distribution with mean $-z_0\sigma$ and variance σ^2, for some constant σ. The percentile method, then, corresponds to the special case of $z_0 = 0$. If this assumption holds, then the endpoints of the interval can be found by

$$G_n^{-1}(\Phi(2z_0 \pm z_{\alpha/2})|\hat{F}n),$$

where Φ is the standard normal cdf, z_α refers to the α^{th} percentile of Φ, and

$$G_n^{-1}(a|\hat{F}n) \equiv \frac{\#\{\theta_i^* \leq a\}}{B}.$$

The bias-correcting constant z_0 is found to be

$$z_0 = \phi^{-1}(G_n(\hat{\theta}|\hat{F}n)).$$

Note that here, as in the percentile method, the transformation w need not be known or estimated. It merely needs to exist.

The BC interval provides correct coverage and approximates the exact limits quite closely in cases where the transformation assumption holds. However, Schenker (1985) notes cases in which the resulting normal distribution is still dependent on the parameter of interest. In response to this, Efron (1987) develops *accelerated, bias-corrected* confidence intervals, BC$_a$. For these intervals, the resulting normal distribution has parameters $-z_0\sigma_\phi$ and σ_ϕ^2, where $\sigma_\phi = \sigma(1 + a\phi)$. The constant a

is referred to as the *acceleration*. The motivation behind this form is that, for BC intervals, the transformation w must both normalize and variance-stabilize, while in BC_a it needs only to normalize. The acceleration constant handles the changing variance. Note that BC intervals correspond to the case $a = 0$.

Hall (1992) recommends a different approach. Recognizing the importance in many cases of bootstrapping pivotal statistics, he recommends computing the bootstrap distribution of statistics of the form,

$$\tau = \frac{\theta - \hat{\theta}}{\hat{\sigma}_\theta}.$$

Then a *percentile-t* interval corresponds to

$$(\hat{\theta} + \hat{\sigma}_\theta \tau^*_{\alpha/2}, \hat{\theta} + \hat{\sigma}_\theta \tau^*_{1-\alpha/2}),$$

where τ^*_α refers to the α^{th} percentile of the bootstrap distribution of τ. Various adjustments (such as stabilization of the variance of $\hat{\theta}$) can be made to this procedure.

Bootstrap confidence intervals have generated much for debate, because it appears that no single approach is best under all circumstances. Since Johnson *et al.* (1982) found biases to be present in proportional hazards estimators, BC intervals are used as the bootstrap intervals in this thesis.

Bootstrapping in regression

In his original paper on the bootstrap, Efron (1979) discusses application of the bootstrap to regression problems. The context is that of a general model with additive errors,

$$Y_i = g_i(\beta) + \epsilon_i, \; i = 1, \ldots, n, \tag{23}$$

there the $\{g_i\}_{i=1}^n$ are known, possibly depending on some p-dimensional nonrandom explanatory variables, $\{\boldsymbol{x}_i\}_{i=1}^n$, and

$$\epsilon_i \overset{iid}{\sim} F,\ i = 1, \ldots, n,$$

for some unknown distribution F centered at zero. Hence, the appropriate resampling scheme is to let $\hat{F}_n$ be the empirical distribution function of

$$\hat{\epsilon}_i = Y_i - g_i(\hat{\boldsymbol{\beta}}),\ i = 1, \ldots, n, \tag{24}$$

for whatever estimator $\hat{\boldsymbol{\beta}}$ is being examined. Then resampling from $\hat{F}_n$ amounts to drawing a new set of residuals, $\{\epsilon_i^*\}_{i=1}^n$ with replacement from $\{\hat{\epsilon}_i\}_{i=1}^n$. Under $\hat{F}_n$ the "true" parameter is $\hat{\boldsymbol{\beta}}$, so the bootstrap sample of responses is created by taking

$$Y_i^* = g_i(\hat{\boldsymbol{\beta}}) + \epsilon_i^*,\ i = 1, \ldots, n.$$

The paper by Wu (1986), with its numerous discussions, provides a great deal of insight into the problems associated with resampling residuals in linear regression. With too few degrees of freedom, the residuals do not resemble a simple random sample from *any* distribution. When the errors are heteroscedastic, the residuals are no longer even exchangeable. The message from Wu and the discussants is that proper modeling of the *entire problem*, not just the mean portion, is crucial to the quality of the bootstrap analysis.

Efron (1982) also discusses a simpler version of the bootstrap for the equation (23). When explanatory variables are present, one could resample vectors $\{(\boldsymbol{x}_i, y_i)^*\}_{i=1}^n$ from $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. The assumption implicit in this procedure is that

$$(\boldsymbol{X}, Y) \sim F^{(p+1)},$$

where $F^{(p+1)}$ is some $p+1$-dimensional distribution function defined on some appropriate support set. Notice that this assumption is *not* compatible with the previous case, where explanatory variables were nonrandom. Freedman (1981) distinguishes between the two cases, calling the first problem the *regression model* and the second problem the *correlation model.*

It is important to realize that in the linear regression model, $\hat{\beta}$ has covariance matrix $\sigma^2(X'X)^{-1}$. This is not the same as the variance under the correlation model, unless ϵ and X are independent (Hall, 1992). Thus, it is clear that employing the wrong resampling scheme will lead to the wrong variance being estimated. Also, there is cause to be concerned about the quality of the correlation-model bootstrap when p is large. Under such circumstances, the resulting empirical cdf is of high dimension, and thus is likely to be an inadequate estimate of the true distribution unless n is extremely large. This "curse of dimensionality" suggests that the use of the correlation-model bootstrap as a simpler approximation to the regression-model bootstrap is to be avoided.

Efron and Gong (1983) use the correlation-model bootstrap in a discriminant analysis for the purpose of stepwise variable selection. They admit that no theory is developed for interpreting the results, but in their problem, four of the nineteen predictors were selected in at least a third of the bootstrap samples, leading them to believe that a relationship existed between these and the response. Efron and Tibshirani (1986) cite several examples in which a bootstrap of complex nonparametric curve estimation techniques is used to suggest a parametric form for a model.

Bootstrap Applications in Survival Analysis

The first application of the bootstrap to censored data appears in Efron (1981a). His goal is to compute a small-sample standard error for the Kaplan-Meier estimator and compare that to the asymptotic estimator, known as Greenwood's formula. He assumes independent random censoring with observed time $T^o = min(T, C)$, where T is the failure time random variable and C is the censor time random variable.

Using Kaplan-Meier techniques to estimate both the failure-time and censor-time distributions, he argues that the proper approach is to resample independently a T^* and a C^* from their respective distributions and take $T^{o*} = min(T^*, C^*)$. However, he shows that this is equivalent to resampling $(T^o, \delta)^*$ as a pair, where δ is the censoring indicator. Using instead this easier approach, he computes bootstrap standard errors for the Channing House data that are quite similar to the corresponding Greenwood values.

Several authors have investigated bootstrapping in proportional hazards models. Chen and George (1985) and Altman and Andersen (1989) each apply the bootstrap to a stepwise variable selection problem, as in Efron and Gong, to data fitting a proportional hazards model. Barlow and Sun (1989) investigate bootstrap confidence intervals for parameters from a model with linear relative risk (constrained to be greater than zero). All of these authors resample vectors $(t^o, \boldsymbol{x}, \delta)$—the observed time, explanatory variables, and censoring indicator—in a manner analogous to the correlation-model bootstrap. Barlow and Sun also resample t^o along with observed likelihood scores and contributions to the information matrix for all individuals. No justification or precedent is offered for this scheme.

Efron and Gong reanalyze the data from Cox (1972) using a bootstrap approach.

Since the explanatory variable has two levels, they treat the data as two separate samples. They then resample separately from the treated and control groups, thus maintaining the original numbers of observations in the two groups. Their results are quite similar to those of Cox, except that they obtain slightly different confidence limits using the percentile method.

Efron and Tibshirani (1986) also analyze the data from Cox, except that they treat the problem as analogous to the correlation model of the previous section. Thus, they resample the explanatory variables along with the times and censoring indicators. They, too, compute a standard error similar to Cox's and compute several different confidence intervals, which differ somewhat from one another.

Both of these analyses fail to capture the essential structure of randomness on the problem. The two-sample method of Efron and Gong would be appropriate if, say, the two groups corresponded to different clinics in which the patients were enrolled. Then resampling patients within clinic would capture the potential for differing populations. Efron and Tibshirani assume that the treatment variable was randomly sampled from some population. In fact, all patients were enrolled in one clinic, and all were assigned to receive their particular treatments. Hence, neither of these resampling schemes is appropriate for this data set.

Karrison attempts to handle a problem with both treatment groups and covariates. Assuming a piecewise exponential (PE) model with I intervals, he proposes a resampling scheme which

- Assumes treatment groups correspond to different populations (as in Efron and Gong);
- Assumes covariates are fixed and correspond to separate multiplicative effects

on the hazard in different populations;

- Allows random censoring to have separate distributions within groups. These may depend on covariates.

He resamples failure times parametrically from his estimated piecewise exponential model for each set of covariates in the group. He also draws a censoring time from the appropriate nonparametric censoring distribution estimate. Taking the minimum of censor and failure times results in observations consisting of a time, a censoring indicator, a group membership indicator, and a covariate vector.

Karrison demonstrates his resampling scheme on two data sets available in the literature. However, both of these have random covariates and fixed (not two-population) treatment effects. He makes no suggestion for how his method could be extended to appropriately handle this situation.

Goals of Papers

It is clear from the preceding review that there is not yet a satisfactory method available with which to handle the problem of bootstrapping in proportional hazards models when explanatory variables are fixed. Also, the estimation of regression parameters in multivariate survival problems has not yet been adequately developed.

In order to address these issues, three papers are presented detailing new advances in methodology. The goals of these papers are, in the order in which they appear:

1. Develop a methodology, analogous to the residual resampling plans in linear regression, for bootstrapping complex regression problems when explanatory

variables are fixed by design. This should be sufficiently flexible as to admit inclusion of complicating factors such as censoring into the solution.

2. Extend this methodology to the problem of multivariate survival analysis. Specifically, using an independence working model for the estimation of regression parameters, develop a resampling plan for the underlying *joint* distribution which will provide appropriate variances and covariances for these estimates.

3. Develop a method for bootstrap estimation in the proportional hazards model which allows resampling from a continuous distribution, which in some cases is completely known up to a parameter value, without full specification of the original distribution of the data.

PAPER I.

A RESIDUAL RESAMPLING PLAN FOR PROPORTIONAL HAZARDS MODELS

ABSTRACT

A resampling plan is introduced for bootstrapping estimators for the Cox (1972) proportional hazards regression model when explanatory variables are nonrandom constants fixed by the design of the experiment. The plan is an analog to the residual-resampling method for regression introduced by Efron (1979). The resampled quantities are a form of generalized residuals and have a distribution that is independent of the explanatory variables. Hence, unlike methods of other authors, this approach does not require resampling of explanatory variables, which is contrary to the assumption that they are nonrandom. An invariance property of the Cox likelihood allows these residuals to be transformed into a convenient scale. Also, the method admits censoring from a class of censoring distributions, which includes the Koziol-Green model. An application to carcinogenesis in rats is discussed.

1. INTRODUCTION

In studying the effects of explanatory variables on failure times, prior knowledge or preliminary analyses often support the assumption that the data come from a distribution featuring the proportional hazards property. The condition of proportional hazards requires that the hazard function for an individual with explanatory variables $\boldsymbol{x}$ can be written as the product of two functions:

$$h(t;\boldsymbol{x}) = h_0(t)g(\boldsymbol{x},\boldsymbol{\beta}). \tag{1.1}$$

Here, t is an observation of T, the positive random variable corresponding to time to failure, $h_0(t)$ is the *baseline hazard function* common to all individuals in the study, and $g(\boldsymbol{x},\boldsymbol{\beta})$ is a positive *relative risk function* of the explanatory variables and some unknown parameters $\boldsymbol{\beta}$. Given an appropriate form for $g(\boldsymbol{x},\boldsymbol{\beta})$, estimation of $\boldsymbol{\beta}$ depends only on the assumptions made about the form of the baseline hazard function.

Cox (1972) considered the case where no assumptions are made about the form of baseline hazard function, and he proposed an estimation procedure based on the likelihood function,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{g(\boldsymbol{x}_i,\boldsymbol{\beta})}{\sum_{l\in\mathcal{R}_i} g(\boldsymbol{x}_l,\boldsymbol{\beta})}. \tag{1.2}$$

Here the product is taken over the k distinct failure times in the data; n is the number

of individuals in the study, and $\mathcal{R}_i$ is the set of indices $\{j : t_j \geq t_{(i)}; j = 1, \ldots, n\}$, called the *risk set* at the i^{th} ordered failure time $t_{(i)}$. This form of the likelihood was later identified to be a *partial likelihood* (Cox, 1975). Kalbfleisch and Prentice (1973) noted that, when $g(\boldsymbol{x}, \boldsymbol{\beta})$ is independent of time and $h_0(t)$ is strictly positive over all open intervals, Cox's likelihood function remains invariant under the group of monotone increasing transformations of t.

Tsiatis (1981), Andersen and Gill (1982), and Bailey (1983) have derived asymptotic results for the estimator $\hat{\boldsymbol{\beta}}$ resulting from maximizing the likelihood (2). They determine, under varying conditions, that $n^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normal with mean zero and variance $I^{-1}(\boldsymbol{\beta})$, the inverse of the Fisher information matrix. This can be consistently estimated by substituting the observed information matrix for the expected information. However, Johnson *et al.* (1982) found that in small samples inferences about $\boldsymbol{\beta}$ can be distorted by the bias exhibited by $\hat{\boldsymbol{\beta}}$, and that the asymptotic variance estimator tends to underestimate the variability of $\hat{\boldsymbol{\beta}}$.

Several authors have proposed using *the bootstrap* for the proportional hazards model. Efron and Gong (1983) and Efron and Tibshirani (1986) used different forms of bootstrap resampling to reanalyze the leukemia remission data set in Cox (1972), obtaining results quite close to those of Cox. Chen and George (1985) and Altman and Andersen (1989) applied the bootstrap to problems of variable selection for the relative risk. In all but one of these papers, explanatory variables were resampled along with times and censoring indicators in a manner similar to the correlation-model bootstrap in Freedman (1981). But resampling the explanatory variables is not always appropriate. If, for example, they represent some fixed treatment groups or dosage levels into which patients are assigned, then these values should be treated

as known constants for each individual, rather that observations sampled from some distribution. Efron and Gong avoided resampling the fixed explanatory variable corresponding to the two treatments in the childhood leukemia data by resampling times and censoring indicators separately within the two treatment groups. The bootstrap regression literature (e.g., Freedman, 1981; Wu, 1986; and Hall, 1992) contains a great deal of discussion on proper conditioning in matching a resampling method to an experiment.

The resampling plan derived in Section 2 provides an analog to the residual-resampling plans used in regression, where explanatory variables are nonrandom, by resampling *generalized residuals*, (Cox and Snell, 1968). Certain types of random censoring can be incorporated naturally into this plan. In Section 3 the carcinogenesis data of Kalbfleisch and Prentice (1980) are analyzed with this new method.

A small-sample study, designed to examine the precision and accuracy of the proposed bootstrap and its counterparts, is described in Section 4. Results of this study are presented in Section 5. Both uncensored and randomly censored data sets are studied, as well as the effects of estimating multiple parameters. The question of sensitivity of this resampling plan to the required censoring forms is discussed briefly.

2. RESIDUAL BOOTSTRAP

In the analysis of failure time data with explanatory variables, a key feature of Cox's likelihood (2) is that it allows for estimation of regression parameters without specification of the form of the baseline hazard function, $h_0(t)$. Thus, assessments of treatment effects, for example, can be obtained without potentially restrictive distributional assumptions.

The residual bootstrap developed in this section is based on the invariance of $L(\boldsymbol{\beta})$ in (2) under the group of monotone increasing transformations of t. Kalbfleisch and Prentice (1973) observed that when explanatory variables do not depend on time, $L(\boldsymbol{\beta})$ depends on $t_1, \ldots, t_n$ only through the risk set $\mathcal{R}_i$; i.e., only through the ranks of the times. The exact form of $h_0(t)$ is unimportant, as long as it is strictly positive over every open interval. Throughout the discussion, it is assumed that $\boldsymbol{x}$ represents some fixed set of constants, which precludes the possibility of time-dependent covariates.

One example of a useful monotone transformation in the analysis of continuous failure time data is the baseline failure time distribution,

$$F_0(t) \equiv Pr(T \leq t).$$

The invariance of $L(\boldsymbol{\beta})$ implies that (2) could equivalently be based on observations of F_0. Data of this type can be said to be in the *probability scale*, as opposed to the

original *time scale.* From (1) the probability of survival beyond time t is

$$S(t;\boldsymbol{x}) = [S_0(t)]^{g(\boldsymbol{x},\boldsymbol{\beta})}, \tag{2.1}$$

for an individual with explanatory variables $\boldsymbol{x}$, and it follows that

$$F_0(t) = 1 - [S(t;\boldsymbol{x})]^{1/g(\boldsymbol{x},\boldsymbol{\beta})}. \tag{2.2}$$

This form for F_0 will be used for generating probability-scale times, needed for the bootstrap analysis of proportional hazards models with nonrandom explanatory variables.

By the probability integral transformation, the distribution of $U \equiv S(T;\boldsymbol{x})$ is Uniform(0,1), regardless of the underlying baseline distribution F_0. Thus,

$$u_i \equiv S(t_i;\boldsymbol{x}_i),\ i = 1,\ldots,n \tag{2.3}$$

are the *iid* variates required for the bootstrap algorithm. These quantities can be viewed as generalized residuals as defined by Cox and Snell (1968). Indeed, these authors explicitly recommend use of the probability integral transformation for creating residuals. Notice, too, that these residuals are closely related to those used by other authors for model diagnostics. The unit-exponential residuals of Crowley and Hu (1977) and Kay (1977) are simply minus-log transforms of the u_i's. We say that the random variable U corresponds to the failure times measured on a *uniform scale.*

As in the additive-error regression analog, residuals must be estimated from the data. For simplicity, temporarily assume that no censoring is present. Estimates of the residuals can be obtained by replacing $S(t,\boldsymbol{x})$ with a consistent estimator in (3). Several candidates can be found, for example, in Kalbfleisch and Prentice (1980). The resulting estimates $\{\hat{u}_i\}_{i=1}^n$ can thus be considered as the population

from which values can be resampled nonparametrically in the bootstrap analysis. For a particular bootstrap sample of residuals, $\{u_i^*\}_{i=1}^n$, probability-scale failure times are computed through (2) as

$$y_i^* = 1 - [u_i^*]^{1/g(\boldsymbol{x},\hat{\boldsymbol{\beta}})},\ i = 1,\ldots,n. \tag{2.4}$$

This plan can be easily adjusted to handle certain types of censoring. Let δ_i be the censoring indicator for the i^{th} individual: $\delta_i = 1$ if individual i is observed to fail, $\delta_i = 0$ if individual i is censored. A residual corresponding to a censored observation can in turn be taken to be a censored observation from its distribution. (Here the censoring is from the left, but that is immaterial to the resampling procedure.) This is commonly done with the unit-exponential residuals (see Crowley and Hu, 1977). Hence, as in Efron (1981a), assigning probability $\frac{1}{n}$ to each of the pairs $\{\hat{u}_i, \delta_i\}_{i=1}^n$ provides a 2-dimensional empirical cdf on $(0,1) \times \{0,1\}$ that can be used to generate bootstrap samples.

Although no assumptions are made about the way in which the residuals and censoring indicators are related, resampling censoring indicators with residuals implies a particular form for the relationship between the failure-time and censoring-time distributions. Specifically, suppose censoring is random, and let C be the random variable corresponding to censoring time with distribution be denoted by H_C. Also, let W correspond to the censoring time in the uniform scale of the failure times; i.e.,

$$W = (1 - F_0^{-1}(C))^{g(\boldsymbol{x},\boldsymbol{\beta})},$$

and denote its distribution by H_W. It is easily shown that taking the pairs (U, δ) as being *iid* is equivalent to assuming that

$$H_C(c) = H_W(S(c; \boldsymbol{x})). \tag{2.5}$$

In other words, the distribution of censoring times must be some function of the conditional distribution of failure times, given $\boldsymbol{x}$.

If the censoring is light (say, 10% or less), this is not likely to be much of a restriction, since there is little information with which to distinguish among possible censoring distributions. Even with heavier censoring, if the censoring fraction is roughly the same for each $\boldsymbol{x}$ and the actual forms of the censoring distributions in the uniform scale do not vary drastically, this resampling plan will generally be adequate. One type of censoring which is likely to cause difficulty, however is type I (fixed endpoint) censoring, particularly when β is far from zero. This problem will be examined in a small simulation study at the end of Section 5.

One model for censoring distributions which satisfies (5) is the well-known *proportional hazards model of random censoring*, sometimes called the Koziol-Green model. Koziol and Green (1976) suggested the relation

$$1 - H_C(c) = (S(c; \boldsymbol{x}))^{\alpha} \tag{2.6}$$

as a possible model for censoring distributions. Using this model, they developed a statistic based on the nonparametric product-limit estimate of the survival distribution for testing goodness-of-fit of parametric survival distributions. Under the Koziol-Green model, the censoring fraction is easily shown to be $\frac{\alpha}{\alpha+1}$. This fact will be used in the simulation study in Section 5.

When censoring is present, it is desirable to use a continuous estimator of $S(t; \boldsymbol{x})$ for estimating residuals. This allows residuals corresponding to censored observations to take on values distinct from those corresponding to failures. Any estimator assuming constant, nonzero hazard between failures will provide this feature; see Breslow (1974), for example.

To summarize the procedure, a nonparametric "residual bootstrap" for the proportional hazards model proceeds as follows:

1. Estimate $\hat{\boldsymbol{\beta}}$ from the original data $(t_i, \boldsymbol{x}_i, \delta_i)$, $i = 1, \ldots, n$, using model (1);

2. Estimate values of $\hat{u}_i = \hat{S}(t_i; \boldsymbol{x}_i)$, $i = 1, \ldots, n$;

3. Resample pairs of observations (u_i^*, δ_i^*), $i = 1, \ldots, n$ from $(\hat{u}_i, \delta_i)$, $i = 1, \ldots, n$ using simple random sampling with replacement;

4. Calculate probability-scale times, $y_i^* = 1 - [u_i^*]^{1/g(\boldsymbol{x}, \hat{\boldsymbol{\beta}})}$, $i = 1, \ldots, n$, where censoring is determined by the corresponding δ_i^*;

5. Obtain $\boldsymbol{\beta}^*$ by maximizing the likelihood

$$L^*(\boldsymbol{\beta}) = \prod_{i=1}^{k^*} \frac{g(\boldsymbol{x}_i, \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}_i^*} g(\boldsymbol{x}_l, \boldsymbol{\beta})};$$

6. Repeat (3)–(5) some large number of times.

It will be seen from the simulations in Section 5 that this procedure does indeed improve both accuracy and precision of estimation of $\hat{\boldsymbol{\beta}}$ and its variance in a variety of situations.

3. CARCINOGENESIS STUDY

The residual bootstrap is illustrated with an analysis of the carcinogenesis data in Kalbfleisch and Prentice (1980). For comparison, the bootstrap method of resampling vectors $(t, \boldsymbol{x}, \delta)$, referred to as the "vector bootstrap," is also applied.

The sample consists of 40 rats exposed to a carcinogen. The rats were randomly assigned into two treatment groups prior to exposure. The primary objective was the evaluation of the relative effectiveness of the two treatment methods in extending lifetime (measured in days) until mortality due to vaginal cancer. The treatment assignment was split 19-21, and there were two censored observations in each group. Since censoring is light, the distribution may reasonably be considered to be of the form (5).

As in Kalbfleisch and Prentice, the loglinear form is used for the relative risk,

$$g(x, \beta) = e^{x\beta},$$

where x takes values 0 or 1 for treatments 1 and 2, respectively. Ties are handled in both analyses using the approximation of Breslow (1974), which also provides the method of calculating residuals used here.

One thousand bootstrap replicates were collected and analyzed for each of the two methods. Histograms of the 1000 bootstrap estimates of β from the two methods are given in Figures 3.1 and 3.2. Both figures exhibit some slight left-skewness,

and the vector bootstrap has a slightly greater dispersion. This is reflected in the estimates of the parameters and their standard errors. The "standard" partial likelihood analysis provides the value $\hat{\beta} = -.60$ for the treatment effect with standard error $\hat{\sigma} = .35$. Estimates from the residual bootstrap were $-.63$ and .37, respectively, while those from the vector bootstrap were $-.62$ and .38, respectively. Incorporating bias correction yields estimates of β with values $-.56$ for the residual bootstrap and $-.57$ for the vector bootstrap. It will be seen in Section 5 that such bias correction improves both the point estimates for β and the corresponding confidence intervals.

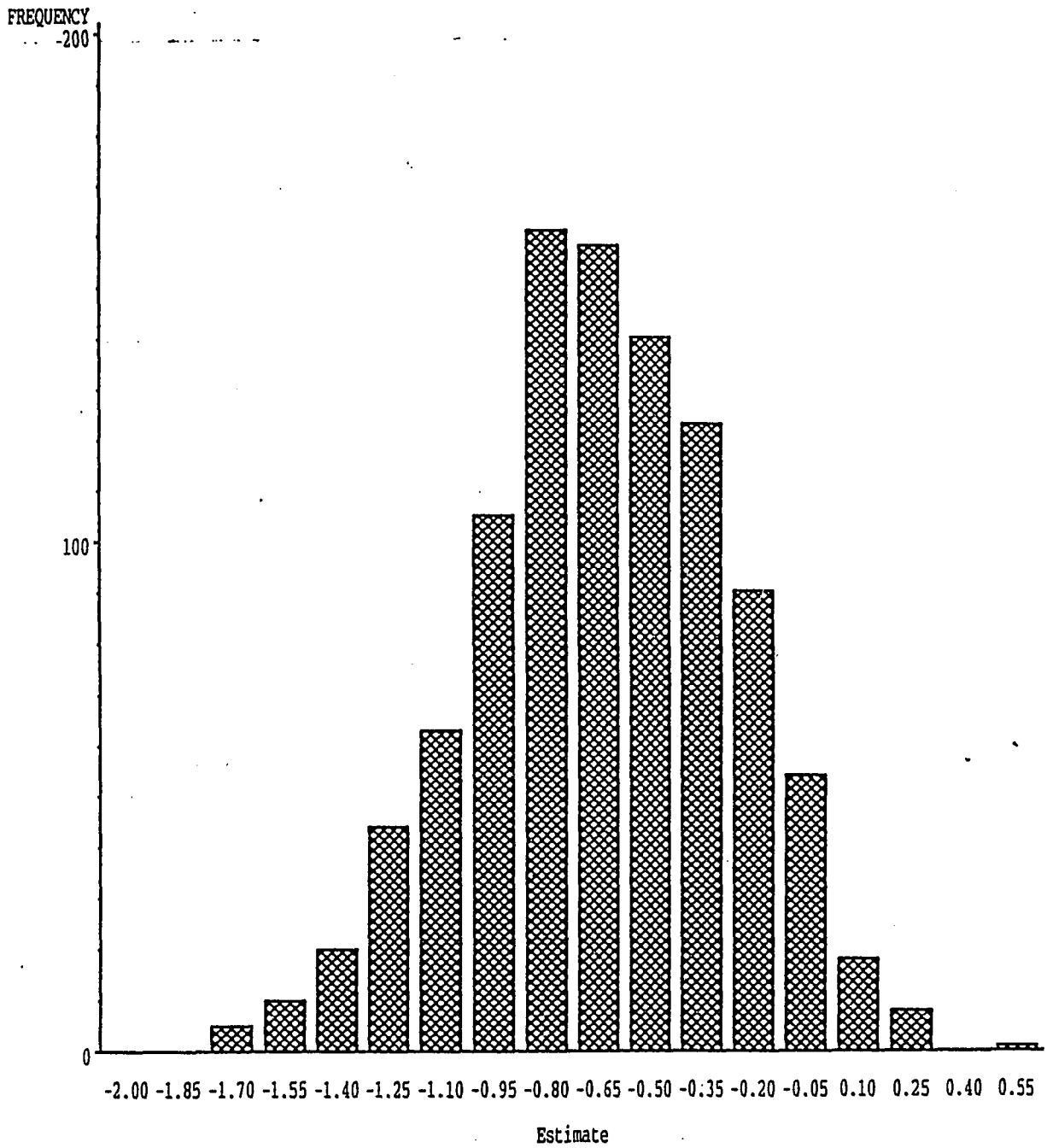

Figure 3.1: Estimated sampling distribution for $\hat{\beta}$ from carcinogenesis data using the residual bootstrap

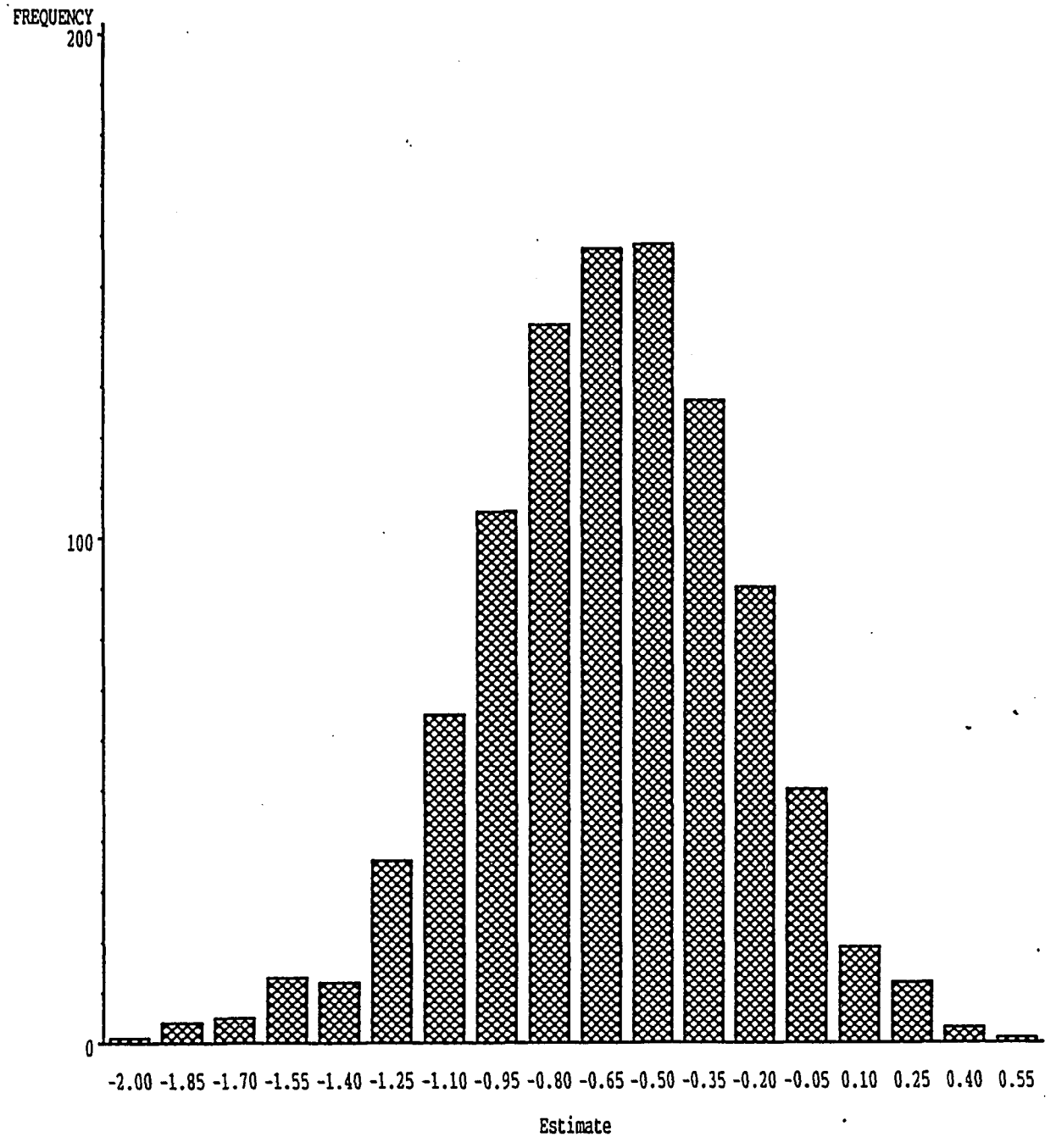

Figure 3.2: Estimated sampling distribution for $\hat{\beta}$ from carcinogenesis data using the vector bootstrap

4. DESCRIPTION OF SIMULATION STUDY

In order to provide an objective basis on which to assess the performance of the proposed procedures, a small Monte Carlo simulation was performed. For more thorough investigation of the estimator resulting from Cox's likelihood, the studies by Johnson *et al.* (1982) and Costanza and Nichola (1982) are recommended. Here a comparison is made of the asymptotic method based on Cox's partial likelihood (referred to in this section as the "standard" method), the residual bootstrap of Section 2, and the "vector bootstrap" as described in the previous section.

Two different estimation situations are considered. In the first situation there is only one explanatory variable, corresponding to two treatment groups of equal size. The parameter value $\beta = 1.5$ is used as motivated by the example in Cox (1972). In the second situation there are five groups of equal size corresponding, for example, to five dosage levels for a particular treatment. The resulting four parameters are assigned values of .2, .4, .6, and .8 (no attempt was made to treat the effect as linear).The relative risk function of the explanatory variable is taken to be $g(\boldsymbol{x}, \boldsymbol{\beta}) = e^{\boldsymbol{x}\boldsymbol{\beta}}$ in each case, with x being a set of either one or four group membership indicators for the two- or five-group situations, respectively.

Samples of sizes 24 or 25 (for the two or five group cases, respectively), 50, and 100 are examined. Both censored and uncensored data are considered, with 20%

censoring provided by the Koziol-Green model. For each of the six combinations of sample sizes and censoring schemes, 1000 Monte Carlo data sets are simulated. The standard method of estimation due to Cox, the residual bootstrap, and the vector bootstrap are applied to each data set. For each bootstrap method, 1000 replicate samples are drawn.

All computations in the simulations and example were performed on DEC workstations at Iowa State University. The programs were written in FORTRAN using double-precision arithmetic. The uniform random number generator proposed by Wichman and Hill (1982) provided the basis for generating data and selecting bootstrap samples. The likelihood maximization algorithm initially uses a modified Newton-Raphson procedure, and switches to the Powell algorithm when the Newton Raphson procedure fails to converge.

For all bootstrap procedures, both raw and bias-corrected estimates are computed. Hence, there are five estimators of each β computed for each sample. These are compared with respect to their average biases and mean squared errors (MSE). The asymptotic variance estimate and the two bootstrap estimates of variance are compared to the estimate of the true finite sample variance, σ^2_{fs}, given by the sample variance of the 1000 Monte Carlo values of $\hat{\beta}$. The MSEs three variance estimates with respect to σ^2_{fs} were used as measures of precision of variance estimation.

Additionally, 80%, 90%, and 95% confidence intervals for β are formed from the various procedures. The usual normal approximation is used to create intervals for the standard Cox estimate and corresponding variance. Bias-corrected percentile limits of Efron (1981b) are also computed for both bootstraps. Since bias correction significantly improves estimation from both bootstrap procedures, only intervals

which include bias correction are considered. Comparisons of coverage percentages and widths were made in all cases. Only results for 90% intervals are presented; results for the other levels are similar.

For the situations involving four parameters, results are generally sufficiently similar across the four parameters to provide a meaningful comparison among methods by averaging parameter estimates, bias values, MSEs or variances. Systematic differences among the parameters are discussed further as needed.

Although the data are generated from a continuous distribution, use of Breslow's approximation to the likelihood (2), as done in this study, can result in negatively-biased estimates of β (Farewell and Prentice, 1980) because both bootstraps impose ties on the resampled data sets. Since multiple selection of the same residual for different values of x results in distinct probability-scale failure times, the effect of ties is expected to be less severe in the residual bootstrap than in the vector bootstrap.

It is important to note that occurrence of monotone likelihood precludes the estimation of β in some samples. As detailed by Bryson and Johnson (1981), β is not estimible if any of the explanatory variables (or certain functions thereof) is monotone with respect to the ordered failure times. In the two-treatment problem, this occurs when the last observed failure for one treatment precedes the first observed failure for the other treatment. This structure will result in a likelihood that has no global maximum for $\beta \in I\!R^p$. As recommended by Bryson and Johnson, such samples are detected and replaced with new simulated samples in the Monte Carlo study. Hence, all estimates of bias and variance are made conditional on the existence of a finite $\hat{\beta}$.

The incidence of monotone likelihood in the Monte Carlo trials and bootstrap replicates is detailed in Table 4.1. As expected, the problem virtually disappears as n

increases, but it is quite prevalent in small samples. In the two-treatment case where the two treatment groups are of equal size, the probability of monotone likelihood for a given sample size n and a given value $\beta \geq 0$ is

$$p_{ML} = \frac{n}{2}e^{-\beta}B(\frac{n}{2}e^{-\beta}, \frac{n}{2}+1) + \frac{n}{2}e^{\beta}B(\frac{n}{2}e^{\beta}, \frac{n}{2}+1),$$

where B(r,s) is the coefficient of the beta distribution. For any practical n, the second term is negligible. For this study, with $n = 24$ and $\beta = 1.5$, $p_{ML} = .0036$, so in 1000 Monte Carlo trials about four are expected to result in data in which $\hat{\beta}$ is too large to estimate. This is not likely to have a large impact on the Cox parameter estimates or asymptotic and finite sample variance estimates.

When studying the performance of bootstrap estimators in simulation studies, however, bootstrap replicates are generated for each Monte Carlo sample. The sample estimate $\hat{\beta}$ corresponds to the "true" β for bootstrap resampling, and larger values of $|\hat{\beta}|$ result in larger proportions of replicates with monotone likelihood. The effect of deleting and replacing these samples is similar to truncating the bootstrap estimate of the distribution of $\hat{\beta}$, which leads to bootstrap estimates of bias and variance that underestimate the true bias and the finite sample variance. The effect this has on bias estimation is clearly seen in Figure 4.1.

Results for $n = 24$ in the one-parameter case are therefore difficult to interpret. Roughly 10% of the Monte Carlo trials resulted in bootstraps with 100 or more cases of monotone likelihood. These generally correspond to the largest values of $\hat{\beta}$. Hence, the simulation results for these cases represent a different set of circumstances than results for larger sample sizes. Nevertheless, means, variances, and confidence intervals are presented for $n = 24$ to illustrate the severity of this sampling-induced bias.

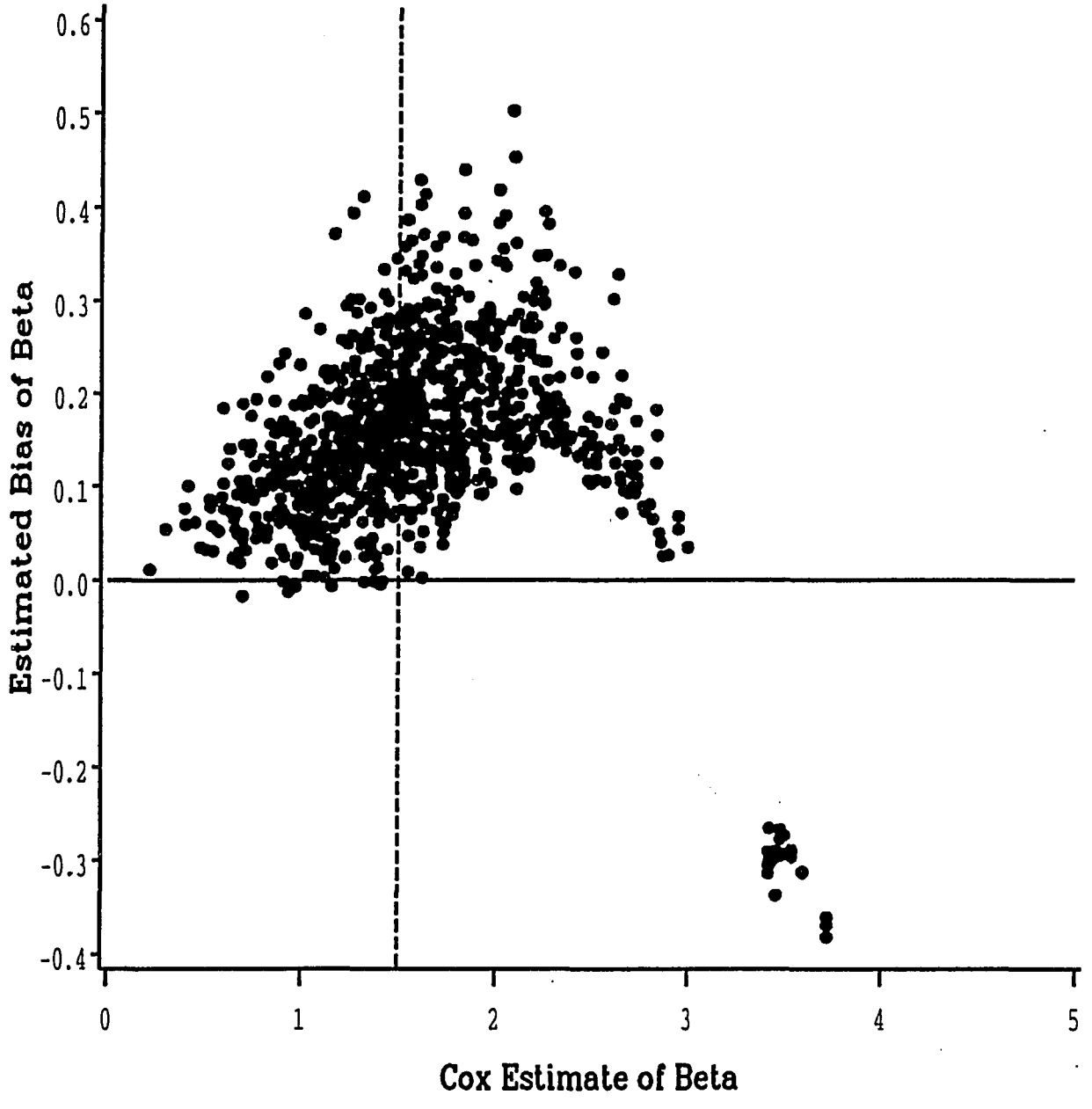

Figure 4.1: Plot of residual bootstrap bias estimates against Cox estimates of β for $n = 24$ and no censoring

In real problems, this phenomenon is of little importance. When values of β are larger than 2, the relative risk exceeds 7, and any reasonable statistical test will detect this. In small samples, large values of β correspond to data sets in which the two groups exhibit very little overlap of failure times. There is little justification for using proportional hazards techniques on such data, and no need for application of a proportional hazards bootstrap. Indeed, an asymptotic variance estimate is of questionable value when a sizable mass of the sampling distribution is placed on infinite values of β. Bootstrap methods provide a useful indication of this situation, even though $\hat{\beta}$ exists for the original data. In larger samples, the probability of a monotone likelihood is minute for any realistic value of β, so again it presents little difficulty. Hence, the recommendation of Bryson and Johnson is valid for practical use.

When there are five treatment groups, the two bootstrap methods show some distinct differences with regard to incidence of monotone likelihood. Because the vector bootstrap allows group sizes to vary in the bootstrap samples, there is a greater possibility of monotone likelihood in small samples. Even when there are no treatment effects, the vector bootstrap can completely omit a treatment group from being resampled. In this study, when $n = 25$, the probability of this is .019. Thus, the vector bootstrap performs poorly for small samples because of the added variability it introduces through resampling the explanatory variables. This is clearly shown in Table 4.1.

For the residual bootstrap, the sampling-induced bias is lower in the four-parameter case than in the one-parameter case because monotonicity of the likelihood is caused by different parameters in different Monte Carlo trials. Thus, the incidence

of monotone likelihood does not correspond as strongly to high values of any single parameter estimate.

Table 4.1: Monotone likelihood cases in simulation study

Number of Params	Censoring	n	M.C. Trials Replaced	Bootstrap	Bootstrap Samples Replaced [a] Mean	Med	Max	%=0 [b]
1	None	24	6	Residual	49	7	1309	17.9
				Vector	49	10	1200	13.3
		50	0	Residual	1	0	180	93.0
				Vector	1	0	138	88.4
		100	0	Residual	0	0	0	100.0
				Vector	0	0	0	100.0
	20%	24	11	Residual	104	24	1509	7.6
				Vector	119	42	1224	4.4
		50	0	Residual	2	0	192	68.2
				Vector	3	0	200	60.3
		100	0	Residual	0	0	1	99.7
				Vector	0	0	1	99.7
4	None	25	0	Residual	28	8	1434	7.2
				Vector	108	83	1428	0.0
		50	0	Residual	0	0	14	97.9
				Vector	1	0	22	63.4
		100	0	Residual	0	0	0	100.0
				Vector	0	0	0	100.0
	20%	25	8	Residual	112	57	1309	0.9
				Vector	425	316	1668	0.0
		50	0	Residual	1	0	107	77.7
				Vector	10	5	322	5.3
		100	0	Residual	0	0	0	100.0
				Vector	0	0	1	99.0

[a] Numbers tabulated over 1000 Monte Carlo trials.

[b] Percentage of trials with no bootstrap replicates replaced.

5. SIMULATION RESULTS

5.1 One Parameter Case

Comparisons of the five estimators of β for the one-parameter case are given in Table 5.1. The standard Cox estimator is clearly seen to have a noticeable upward bias, particularly for the smallest sample size. Both bootstrap methods reflect this with raw estimates that are even farther from $\beta = 1.5$. The residual bootstrap tends to overestimate the bias, resulting in bias-corrected estimates that are lower than the true parameter value. The bias-corrected vector bootstrap estimates appear to be reasonably accurate for $n \geq 50$. Neither bootstrap fares well at $n = 24$, due largely to the problems with monotone likelihood.

The bias-corrected residual bootstrap provides roughly a 10% improvement in MSE over the standard method for $n \geq 50$. Improvements of about half that size are obtained with the bias-corrected vector bootstrap. In these same cases, using raw estimates of either bootstrap causes an increase in MSE of about 5–30%. Thus, bias correction provides a distinct improvement in bootstrap estimation of β and should be used whenever bootstrap methods are applied to this problem.

Estimates of $var(\hat{\beta})$ for the three estimation methods are given in Table 5.2. The comparisons show that the inverse of the information consistently underestimates the finite sample variance, by as much as 12% in the smallest sample size. As with

Table 5.1: Simulation results: Biases and mean squared errors from estimation of β in the one parameter case

Censor	Estimation Method	$n = 24$		$n = 50$		$n = 100$	
		Bias	MSE	Bias	MSE	Bias	MSE
None	Standard	.094	.376	.031	.138	.039	.071
	Resid-BC [a]	-.057	.399	-.057	.124	-.011	.065
	Vec-BC [b]	.043	.451	-.017	.131	.017	.068
	Resid-Raw [c]	.246	.424	.118	.175	.079	.083
	Vec-Raw [d]	.146	.333	.078	.154	.061	.076
20%	Standard	.072	.444	.043	.175	.037	.084
	Resid-BC	-.044	.495	-.051	.153	-.005	.078
	Vec-BC	.058	.567	-.011	.162	.013	.080
	Resid-Raw	.188	.449	.137	.224	.080	.098
	Vec-Raw	.086	.357	.098	.197	.062	.090

[a]Bias-corrected residual bootstrap.
[b]Bias-corrected vector bootstrap.
[c]Residual bootstrap, no bias correction.
[d]Vector bootstrap, no bias correction.

the estimation of β, the MSEs of the asymptotic variance estimator go down with increasing n and are larger when 20% random censoring is present.

The bootstraps, on the other hand, show inconsistent results. At $n = 24$, both bootstraps underestimate $var(\hat{\beta})$ on the average, but they have noticeably smaller MSEs than the asymptotic estimator. This is largely a result of the deflation effect of the monotone likelihood—the potentially largest variances, corresponding to large values $\hat{\beta}$, are underrepresented. For the two larger sample sizes, the vector bootstrap estimate of $var(\hat{\beta})$ is larger than that of the residual bootstrap. This is to

Table 5.2: Simulation results: Sampling variance estimates with mean squared errors in the one parameter case

		$n = 24$		$n = 50$		$n = 100$	
Censor	Estimation Method	$\hat{\text{Var}}(\hat{\beta})$	MSE[a]	$\hat{\text{Var}}(\hat{\beta})$	MSE	$\hat{\text{Var}}(\hat{\beta})$	MSE
None	FSV [b]	.367	—	.138	—	.070	—
	Standard [c]	.324	2975	.133	115	.063	13
	Res. Boot [d]	.328	664	.150	210	.066	18
	Vec. Boot [e]	.348	763	.162	384	.069	28
20%	FSV	.439	—	.173	—	.083	—
	Standard	.409	4216	.168	188	.080	19
	Res. Boot	.359	1193	.195	393	.083	31
	Vec. Boot	.379	1168	.207	559	.087	45

[a] All MSEs are $\times 10^{-5}$.
[b] Variance of the 1000 Monte Carlo estimates of β.
[c] Asymptotic estimator.
[d] Residual Bootstrap.
[e] Vector Bootstrap.

be expected, since the latter method does not include variability due to resampling of x. Also, the MSEs of both bootstrap variance estimates are larger than that of the standard estimator, but all of the values are quite small relative to the variances being estimated.

The confidence interval results in Table 5.3 shows that, except for the smallest sample size where interpretation of bootstrap results is difficult, all confidence intervals provided reasonably good coverage at the 90% level. The vector bootstrap intervals are the widest among the three, but they offer no noteworthy gain in cov-

Table 5.3: Simulation results: 90% confidence interval widths and coverages in the one parameter case

		$n = 24$		$n = 50$		$n = 100$	
Censor	Interval Method	Width	Coverage	Width	Coverage	Width	Coverage
None	Normal[a]	1.83	.903	1.19	.911	0.83	.884
	Resid BC [b]	1.73	.858	1.19	.899	0.82	.881
	Vec BC [c]	1.87	.877	1.26	.912	0.85	.879
20%	Normal	2.05	.906	1.34	.903	0.93	.907
	Resid BC	1.88	.849	1.35	.893	0.93	.902
	Vec BC	1.96	.855	1.43	.891	0.95	.904

[a]Based on asymptotic normal approximation with standard estimators of β and $\text{Var}(\hat{\beta})$.

[b]Bias-corrected percentile intervals for the residual bootstrap.

[c]Bias-corrected percentile intervals for the vector bootstrap.

erage.

5.2 Four Parameter Case

Results from the estimation of the four-dimensional vector β are presented in Table 5.4. The average biases of the standard estimator over the four parameters are somewhat larger for $n \leq 50$ than they were in the one-parameter case. The four parameter estimates (not displayed) show a general increase in bias with the parameter value. Average MSEs for all standard estimates are roughly twice as large as those obtained for the one parameter case.

The bias-corrected bootstraps seem to provide some bias reduction for all sample

Table 5.4: Simulation results: Biases and mean squared errors from estimation of β in the four parameter case

Censor	Estimation Method	$n = 25$		$n = 50$		$n = 100$	
		Bias	MSE	Bias	MSE	Bias	MSE
None	Standard	.111	.791	.043	.278	.035	.126
	Resid-BC [a]	-.043	.475	-.035	.204	-.005	.108
	Vec-BC [b]	.049	.689	.000	.236	.013	.116
	Resid-Raw [c]	.265	1.273	.121	.379	.075	.150
	Vec-Raw [d]	.173	.966	.086	.334	.057	.139
20%	Standard	.114	.920	.066	.347	.018	.149
	Resid-BC	-.025	.599	-.016	.251	-.019	.129
	Vec-BC	.071	.926	.016	.296	-.004	.138
	Resid-Raw	.226	1.388	.149	.478	.055	.174
	Vec-Raw	.156	.996	.116	.417	.040	.163

[a]Bias-corrected residual bootstrap.
[b]Bias-corrected vector bootstrap.
[c]Residual bootstrap, no bias correction.
[d]Vector bootstrap, no bias correction.

sizes, although the vector bootstrap results, in particular, are difficult to interpret at $n = 25$ due to high incidence of monotone likelihood. The residual bootstrap provides a substantial decrease in MSE, ranging from about 14% in the largest samples to 30–40% in the smallest. The vector bootstrap tends to provide reductions that are about half as large. The two raw bootstrap estimators uniformly provide the largest biases and MSEs for all censoring-by-sample size combinations.

Estimation of $var(\hat{\beta})$ is also more difficult in the multiparameter problem. Table 5.5 shows that the average finite sample variances and the average MSEs over the

Table 5.5: Simulation results: Sampling variance estimates with mean squared errors in the four parameter case

		$n = 25$		$n = 50$		$n = 100$	
Censor	Estimation Method	$\hat{Var}(\hat{\beta})$	MSE [a]	$\hat{Var}(\hat{\beta})$	MSE	$\hat{Var}(\hat{\beta})$	MSE
None	FSV [b]	.775	—	.275	—	.125	—
	Standard [c]	.511	8431	.228	285	.109	23
	Res. Boot [d]	.769	1068	.288	157	.122	8
	Vec. Boot [e]	1.057	24714	.352	1804	.132	90
20%	FSV	.904	—	.342	—	.148	—
	Standard	.763	10382	.294	519	.137	32
	Res. Boot	.912	2430	.370	446	.153	24
	Vec. Boot	1.093	20489	.444	2939	.166	166

[a] All MSEs are $\times 10^{-5}$.
[b] Variance of the 1000 Monte Carlo estimates of β.
[c] Asymptotic estimator.
[d] Residual Bootstrap.
[e] Vector Bootstrap.

four parameters represent large increases from the one-parameter case. The asymptotic variance estimator underestimates the finite sample variance in each of the situations. In the smallest sample size, this estimate averages only about 70% of the finite sample variance.

The residual bootstrap provides substantial improvement in estimation of $var(\hat{\beta})$ over the standard method. Residual bootstrap estimates of variance tend to be quite close to the desired values, and the MSEs range from 15% to 75% better than those of the asymptotic estimator. The vector bootstrap, on the other hand, yields a 6–36%

Table 5.6: Simulation results: 90% confidence interval widths and coverage in the four parameter case

		$n = 25$		$n = 50$		$n = 100$	
Censor	Interval Method	Width	Coverage	Width	Coverage	Width	Coverage
None	Normal [a]	2.34	.846	1.57	.878	1.09	.882
	Resid BC [b]	2.80	.929	1.73	.933	1.14	.908
	Vec BC [c]	3.28	.905	1.88	.916	1.18	.901
20%	Normal	2.66	.858	1.78	.884	1.22	.892
	Resid BC	3.06	.913	1.95	.926	1.28	.912
	Vec BC	3.32	.870	2.12	.923	1.32	.909

[a]Based on asymptotic normal approximation with standard estimators of β and $\text{Var}(\hat{\beta})$.

[b]Bias-corrected percentile intervals for the residual bootstrap.

[c]Bias-corrected percentile intervals for the vector bootstrap.

overestimate of the finite sample variance, and it has MSEs two to six times larger than the asymptotic estimator.

The 90% confidence interval results presented in Table 5.6 tell a similar story. The standard normal intervals are too short, providing inadequate coverage. The residual bootstrap intervals are longer and slightly conservative. Intervals based on the vector bootstrap are the longest of the three, but they provide no greater coverage than their residual bootstrap counterparts.

5.3 Sensitivity to Different Censoring Schemes

In order to assess the sensitivity of the residual bootstrap to deviations from the assumed censoring scheme, an additional series of simulations was run using 20% type I (fixed endpoint) censoring. When the parameters are nonzero, censoring in this manner results in residuals with censoring at a different point in the uniform scale for each distinct $\boldsymbol{x}$. Hence, the censoring pattern of resampled data does not resemble that of the original data.

This affects the performance of the residual bootstrap, but the effect seems to decrease with increasing numbers of parameters. When only one parameter is being estimated, residual bootstrap estimates of that parameter are biased slightly more toward zero than before, although their MSEs are still better than those of the standard method. Residual bootstrap estimates of $var(\hat{\boldsymbol{\beta}})$ are 30–45% too high, however, resulting in confidence intervals that are too wide and have supranominal coverage.

When four parameter estimates are required, the average biases and MSEs are comparable to those obtained under the other two censoring plans. Variance estimates for $\hat{\boldsymbol{\beta}}$ are conservative, but not as much so as those from the vector bootstrap. Mean squared errors for variance estimation are about 50% higher for the residual bootstrap than for the asymptotic estimator (which is nearer the finite sample variance than in the other censoring types), but they are still quite a bit smaller than those of the vector bootstrap. The pattern of widths and coverages for confidence intervals is similar to the other four-parameter cases, although the standard normal intervals do an adequate job here for $n \geq 50$.

These findings can be expected to be more extreme with greater censoring frac-

tions. However, there are many other censoring patterns for which the residual bootstrap may be expected to perform at least as well as the vector bootstrap when explanatory variables are fixed. Further study is needed to clarify the general robustness of this method.

6. DISCUSSION

The bias-corrected residual bootstrap performs well for estimating regression parameters in proportional hazards models where explanatory variables are fixed by the experimental design. Its advantages are especially evident in the simulation results for the four-parameter model, where it consistently provides estimates of regression parameters and sampling variances with the smallest mean squared errors among the methods compared. Variance estimates obtained from the standard asymptotic approach tend to be too small, and corresponding confidence intervals do not always provide adequate coverage. Bias-corrected percentile confidence intervals for the residual bootstrap tend to provide more than the nominal coverage, but they are narrower than the corresponding intervals obtained from the vector bootstrap. The bootstrap methods also provide a valuable indication of the reliability of the proportional hazards assumption through their assessment of the incidence of monotone likelihood problems near the estimated parameter value.

One issue that has not been addressed is the quality of the estimator of the residuals that are used in the resampling plan. Lagakos (1980) expresses some concern regarding the small-sample properties of this method of estimating residuals. Crowley and Storer (1983) present a simulation study which indicates that the average values of residuals computed in this manner are nearly the same as their expected values, but

that the variances tend to be somewhat smaller. Unpublished simulations performed in conjunction with the present work indicate that the biases of the order statistics of the residuals approach zero at a rate of roughly n^{-1}. Other available estimators of this function (see Kalbfleisch and Prentice, 1980) might yield some improvement. Alternatively, modifications of generalized residuals proposed by Cox and Snell (1968) to give the residuals properties more representative of their true distributions may also be useful. Since the computation of residuals is central to the resampling method, further research is needed to investigate the properties of their various estimators in the present context.

Also, the question of the best bootstrap confidence interval for this problem remains open. Bias-corrected percentile methods of Efron (1981b) were the best among those presented here, but other candidates remain. Efron (1987) has developed "accelerated, bias-corrected" intervals which require the estimation of an additional parameter but may nonetheless be of some use here. Perhaps a better candidate is the percentile-t interval (see Hall, 1992), which possesses some favorable properties when the estimator is known to be asymptotically normal.

Tied data also present an obstacle, as mentioned in Section 4. All simulation results given here were based on estimation in continuous models, where the original data contain no tied failure times. Properties of the bootstrap estimators are not known when the original data contains tied failure times. Furthermore, resampling induces ties, which prevents direct application of large-sample theoretical results for continuous proportional hazards models (e.g., Tsiatis, 1981; Andersen and Gill, 1982; Bailey, 1983) to bootstrap estimation. The extent to which ties induced by resampling can affect the resulting bootstrap estimates in small samples is not entirely clear. In

this regard, the residual bootstrap has some advantage over the vector bootstrap in that it induces somewhat fewer ties in the bootstrap replicates.

Finally, there is potential for interesting extensions of the residual bootstrap. Work is underway to investigate its use in multivariate survival problems. In some multivariate problems, marginal proportional hazards models may be appropriate. In such cases, a well-designed bootstrap could provide both bias-corrected parameter estimates and estimates of the finite sample variance, including cross-marginal covariances, all without specification of a model for association. More generally, residual bootstrap methods can be extended to many other situations where generalized residuals can be identified.

BIBLIOGRAPHY

Altman, D. G. and Andersen, P. K. (1989). Bootstrap Investigation of the Stability of the Cox Regression Model. *Statistics in Medicine*, **8**, 771–783.

Andersen, P. K. and Gill, R.D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Approach. *Annals of Statistics*, **10**, 1100–1120.

Bailey, K. R. (1983). The Asymptotic Joint Distribution of Regression and Survival Parameter Estimates in the Cox Model. *Annals of Statistics*, **11**, 39–48.

Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, **30**, 89–99.

Bryson, M. C. and Johnson, M. E. (1981). The Incidence of Monotone Likelihood in the Cox Model. *Technometrics*, **23**, 381–383.

Chen, C. H. and George, S. L. (1985). The Bootstrap and Identification of Prognostic Factors via Cox's Proportional Hazards Regression Model.

Statistics in Medicine, **4**, 39–46.

Costanza, M. C. and Nichola, P. S. (1982). Effect of Random Censoring on Cox-Breslow Survival Methods: A Simulation Study. *ASA Proceedings from the Section on Statistical Computation*, 258–262.

Cox, D. R. (1972). Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–202.

Cox, D. R. (1975). Partial Likelihood. *Biometrika*, **62**, 269–276.

Cox, D. R. and Snell, E. J. (1968). A General Definition of Residuals (with discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 248–265.

Crowley, J. and Hu, M. (1977). Covariance Analysis of Heart Transplant Survival Data. *Journal of the American Statistical Association*, **77**, 27–36.

Crowley, J. and Storer, B. E. (1983). Comment on "A Reanalysis of the Stanford Heart Transplant Data." *Journal of the American Statistical Association*, **78**, 277–281.

Efron, B. (1981a). Censored Data and the Bootstrap. *Journal of the American Statistical Association*, **76**, 312–319.

Efron, B. (1981b). Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics*, **9**, 139–172.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. *SIAM monograph*, **38**, CBMS-NSF.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, **82**, 171–185.

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, Jackknife, and Crossvalidation. *American Statistician*, **37**, 36–48.

Efron, B. and Tibshirani, R. J. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, **1**, 54–77.

Farewell, V. T. and Prentice, R. L. (1980). The Approximation of Partial Likelihood with Emphasis on Case Control Studies. *Biometrika*, **67**, 273–278.

Freedman, D. A. (1981). Bootstrap Regression Models. *Annals of Statistics*, **9**, 1218–1228.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer-Verlag.

Johnson, M. E., Tolley, H. D., Bryson, M. C., and Goldman, A. S. (1982). Covariate Analysis of Survival Data: A Small Sample Study of Cox's Model. *Biometrics*, **38** 685–698.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal Likelihoods Based of Cox's Regression and Life Model. *Biometrika*, **60**, 267–278.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Kay, R. (1977). Proportional Hazard Regression Models and the Analysis of Censored Survival Data. *Applied Statistics*, **26**, 227–237.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises Statistic for Randomly Censored Data. *Biometrika*, **63**, 465–474.

Lagakos S. W. (1980). The Graphical Evaluation of Explanatory Variables in Proportional Hazards Regression Models. *Biometrika*, **68**, 93–98.

Tsiatis, A. A. (1981). A Large Sample Study of Cox's Regression Model. *Annals of Statistics*, **9**, 93–108.

Wichman, B. A. and Hill, I. D. (1982). An efficient and portable pseudo-random number generator. *Applied Statistics*, **31**, 188–190.

Wu, C. F. J. (1986). Jackknife, Bootstrap, and Other Resampling Methods in Regression. *Annals of Statistics*, **14**, 1261–1295.

PAPER II.

BOOTSTRAPPING IN MULTIVARIATE SURVIVAL ANALYSIS

ABSTRACT

Bootstrap methods are proposed for estimating sampling distributions and statistics for regression parameters in multivariate survival data. We use an *Independence Working Model* (IWM) approach, fitting margins independently, to obtain consistent estimates of the parameters in the marginal models. Resampling procedures, however, are applied to an appropriate *joint* distribution to estimate covariance matrices, make bias corrections, and construct confidence intervals. The proposed methods allow for fixed or random explanatory variables, using extensions of existing resampling schemes (Loughin, 1993) and they permit the possibility of random censoring. An application is shown for the viral positivity time data previously analyzed by Wei, Lin, and Weissfeld (1989). A simulation study of small-sample properties shows that the proposed bootstrap procedures provide substantial improvements over the robust variance estimator proposed by Wei, Lin, and Weissfeld (1989).

1. INTRODUCTION

In studies of event times, it is often the case that more than one event can occur to each study subject. This is seen in engineering applications when a system has several distinct components, each of which is subject to failure. In health studies, time until onset of different diseases or symptoms may be of interest. Wei, Lin, and Weissfeld (1989) describe another type of multivariate problem: a repeated measures study in which monthly blood samples were drawn from each patient and studied until some marker for viral positivity was attained. Discussions of multivariate survival problems are presented, for example, in Kalbfleisch and Prentice (1980).

In survival studies, it is generally of interest to compare the relative effects of one or more explanatory variables, such as treatment assignments, on the distribution of these event times. The univariate survival problem has been well studied, but less attention has been given to multivariate data. A variety of parametric models have been proposed (see Hougaard, 1987, for a review), but the structure they assume for the association among failure times is often complex and difficult to verify in real data. For example, in the distribution put forth by Crowder (1989), the null hypothesis of independence of failure times is not easily tested, since the corresponding parameterization renders another parameter nonidentifiable.

Robust estimation approaches adopted by Huster, Brookmeyer, and Self (1989)

and Wei, Lin, and Weissfeld (1989) avoid the specification of any form for this association. Instead, an *Independence Working Model* (IWM) is used, in which models are formulated independently for each failure type. Despite the potentially incorrect assumption of independence, consistent parameter estimates are obtained if the marginal models are correctly specified. Both sets of authors then propose related "robust" estimators of the covariance matrix of the estimated regression parameters.

In Section 2 a bootstrap application to the IWM is devised in which data are resampled from the underlying *joint* empirical distribution. The method not only provides an estimate of the covariance matrix, but it can also reduce biases in parameter estimates. The method is developed further in Section 3 for the case in which explanatory variables are fixed by design, as would be the situation when patients are assigned a treatment regimen upon entry into a study. Generalized residuals of Cox and Snell (1968) are used to create *iid* quantities and estimates of these quantities are used for resampling. The result is a multivariate extension of the residual bootstrap of Loughin (1993).

In Section 4 the repeated measures data presented by Wei *et al.* are reanalyzed with this new method. The simulation study presented in Section 5 and Section 6 shows that the residual bootstrap provides substantial improvements over the robust estimators under a variety of circumstances. Extensions of these techniques are discussed in Section 7.

2. BOOTSTRAPPING IN THE INDEPENDENCE WORKING MODEL

In separate work, Huster, Brookmeyer, and Self (1989) and Wei, Lin, and Weissfeld (1989) propose Independence Working Models for multivariate survival data. Huster *et al.* propose modeling paired data (disease status in two eyes) with independent Weibull distributions, while Wei *et al.* use marginal proportional hazards models on the viral positivity data which are analyzed Section 4. Certainly, other marginal models could also be used with the IWM approach.

The general appeal of the IWM formulation is that marginal forms can be taken to be familiar, well-studied models. Goodness-of-fit and omitted-variable tests can be done within margins, allowing the practitioner great flexibility in developing models. The complexity of carrying out these procedures for the multivariate problem is thus avoided, but the trade-off is a potential loss of efficiency for not making full use of information on the joint distribution of survival times.

To formalize ideas, suppose M different event times are collected on each of n subjects. Let

$$\boldsymbol{T} = (T_1, \ldots, T_M)$$

where T_m is the event time random variable for the m^{th} event type. Each subject is assumed to be at risk to each event type. Furthermore, suppose values for a

corresponding set of explanatory variables,

$$\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M),$$

are obtained for each subject, where $\boldsymbol{x}_m$ is a $1 \times p_m$ vector. These may be the same for all margins, or they may differ arbitrarily. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$ be a set of regression parameters for $\boldsymbol{x}$, where $\boldsymbol{\beta}_m$ is a $q_m \times 1$ vector. Define the *joint survivor function* by

$$S(\boldsymbol{t}; \boldsymbol{x}, \boldsymbol{\beta}) \equiv Pr\{T_1 > t_1, \ldots, T_M > t_M; \boldsymbol{x}, \boldsymbol{\beta}\}, \tag{2.1}$$

and the *marginal survivor functions* by

$$S_m(t_m; \boldsymbol{x}_m, \boldsymbol{\beta}_m) \equiv Pr\{T_m > t_m; \boldsymbol{x}_m, \boldsymbol{\beta}_m\} \; m = 1, \ldots, M.$$

Suppose, as well, that, corresponding to the joint and marginal survivor functions, there are density functions, $f(\boldsymbol{t}; \boldsymbol{x}, \boldsymbol{\beta})$ and $f_m(t_m; \boldsymbol{x}_m, \boldsymbol{\beta})$, respectively. For ease of presentation, we restrict attention in the following to two dimensions; extensions to higher dimensions are immediate.

Let $\boldsymbol{t}^{o(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{\delta}^{(i)}$ be the observed times, explanatory variables, and censoring indicators for the i^{th} individual, where $\delta_m^{(i)} = 1$ if $t_m^{o(i)}$ is an observed failure, and $\delta_m^{(i)} = 0$ if the observation in that margin is censored at $t_m^{o(i)}$. Then a full likelihood for $\boldsymbol{\beta}$ based on these data is

$$\begin{aligned} L(\boldsymbol{\beta}) \;=\; & \prod_{i=1}^{n} \left[f(\boldsymbol{t}^{o(i)}; \boldsymbol{x}^{(i)}, \boldsymbol{\beta}) \right]^{\delta_1^{(i)} \delta_2^{(i)}} \times \left[S(\boldsymbol{t}^{o(i)}; \boldsymbol{x}^{(i)}, \boldsymbol{\beta}) \right]^{(1-\delta_1^{(i)})(1-\delta_2^{(i)})} \\ & \times \left[\frac{\partial S(\boldsymbol{t}; \boldsymbol{x}^{(i)}, \boldsymbol{\beta})}{\partial t_1} \Big|_{\boldsymbol{t}=(t_1^{o,(i)}, t_2^{o,(i)})} \right]^{(1-\delta_1^{(i)})\delta_2^{(i)}} \end{aligned}$$

$$\times \left[\frac{\partial S(t; x^{(i)}, \beta)}{\partial t_2} \Big|_{t=(t_1^{o,(i)}, t_2^{o,(i)})} \right]^{\delta_1^{(i)}(1-\delta_2^{(i)})} . \tag{2.2}$$

An estimate $\tilde{\beta}$ of β is found by maximizing (2) for the given data.

The IWM uses only marginal quantities in formulating the likelihood:

$$L_{IWM}(\beta) = \prod_{i=1}^{n} \prod_{m=1}^{2} \left[f_m(t_m^{o(i)}; x_m^{(i)}, \beta_m) \right]^{\delta_m^{(i)}} \left[S_m(t_m^{o(i)}; x_m^{(i)}, \beta_m) \right]^{(1-\delta_m^{(i)})} . \tag{2.3}$$

Maximization of $L_{IWM}(\beta)$ results in an IWM estimate $\hat{\beta}$. Huster *et al.* and Wei *et al.* show that $\hat{\beta}$ is consistent for β under mild regularity conditions as long as the marginal distributions are correctly specified. However, inverting the resulting information matrix will not generally yield a consistent estimate of the sampling variance, $V(\hat{\beta})$.

It is possible to obtain a consistent estimate of this covariance matrix using a robust estimation result from Huber (1967) (see also Royall, 1986). Define

$$A^{(i)}(\beta) \equiv \prod_{m=1}^{2} \left[f_m(t_m^{o(i)}; x_m^{(i)}, \beta_m) \right]^{\delta_m^{(i)}} \left[S_m(t_m^{o(i)}; x_m^{(i)}, \beta_m) \right]^{(1-\delta_m^{(i)})}$$

$i = 1, \ldots, n$ to be the i^{th} term in the likelihood (3). Define the individual score by

$$U^{(i)}(\beta) \equiv \frac{\partial}{\partial \beta} log A^{(i)}(\beta)$$

and the observed information matrix by

$$I(\beta) \equiv -\frac{\partial^2}{\partial \beta \partial \beta'} log L_{IWM}(\beta).$$

Then, under some mild regularity conditions on $f(t; x, \beta)$, application of Huber's results provides that the asymptotic variance of the IWM estimator $\hat{\beta}$ is consistently

estimated by the "sandwich estimator,"

$$\hat{V}_A(\hat{\beta}) = I^{-1}(\hat{\beta})\Lambda(\hat{\beta})I^{-1}(\hat{\beta}), \tag{2.4}$$

where

$$\Lambda(\beta) = \sum_{i=1}^{n} U^{(i)}(\beta)[U^{(i)}(\beta)]'.$$

When the independence assumption holds, $\hat{V}_A$ reduces to $I^{-1}(\hat{\beta})$, the asymptotic estimator under the IWM. However, asymptotic variance estimates do not always accurately reflect the small-sample variability of $\hat{\beta}$. Simulation studies by Johnson *et al.* (1982) and Loughin (1993) suggest that in the univariate problem the asymptotic estimate of $V(\hat{\beta})$ tends to be biased low in small samples, particularly when β is of more than one dimension. Bootstrap methods can substantially improve the estimation of $V(\hat{\beta})$ in small samples.

The idea behind the bootstrap (Efron, 1979) is to approximate $\mathcal{G}_n(\hat{\beta}|\mathcal{F}_\beta)$, the sampling distribution of an estimator $\hat{\beta}$ based on a simple random sample of size n from $\mathcal{F}_\beta$, with $\mathcal{G}_n(\beta^*|\hat{F}_n)$, where $\hat{\mathcal{F}}_n$ is an approximation to the true distribution $\mathcal{F}_\beta$ based on the original sample (e.g., the empirical df or a parametric form evaluated at $\hat{\beta}$), and β^* is an estimate of β based on a simple random sample from $\hat{\mathcal{F}}_n$. In survival analysis applications, direct analytical evaluation of the properties of $\mathcal{G}_n(\beta^*|\hat{F}_n)$ is generally not feasible, and so information about $\mathcal{G}_n(\beta^*|\hat{F}_n)$ is obtained through repeated Monte Carlo resampling from $\hat{\mathcal{F}}_n$.

Consider in this context the problem of regression parameter estimation with multivariate survival data. Suppose for the moment that there is no censoring. It might be reasonable to assume that $(\boldsymbol{t}^{(i)}, \boldsymbol{x}^{(i)})$ are drawn from some unknown $(2+P)$-dimensional distribution $\mathcal{F}$, for $P = p_1 + p_2$. Then $S(\boldsymbol{t}; \boldsymbol{x}, \beta)$ is the joint

survivor function of $\boldsymbol{T}$ conditional on $\boldsymbol{x}$, and $S_m(t_m; \boldsymbol{x}_m, \boldsymbol{\beta}_m)$ are the marginal survivor functions of $\boldsymbol{T}_m$ conditional of $\boldsymbol{x}_m$, $m = 1, 2$.

It is not possible to make use of the (assumed known) forms for S_m in a resampling procedure without some knowledge of the joint survivor function and the marginal distribution of $\boldsymbol{x}$. Alternatively, nonparametric resampling can be used. Let $\hat{\mathcal{F}}_n$ be the empirical distribution function of $(T_1, T_2, \boldsymbol{x}_1, \boldsymbol{x}_2)$. Let $\boldsymbol{\beta}^{*b}$, $b = 1, \ldots, B$ be the value of $\boldsymbol{\beta}$ maximizing (3) based on the b^{th} random sample of size n from $\hat{\mathcal{F}}_n$, for some large B. Then we can define

$$\bar{\boldsymbol{\beta}}^*_{nB} \equiv \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{\beta}^{*b} \tag{2.5}$$

and

$$V^*_{nB}(\hat{\boldsymbol{\beta}}) \equiv \frac{1}{B-1} \sum_{b=1}^{B} \left(\boldsymbol{\beta}^{*b} - \bar{\boldsymbol{\beta}}^*_{nB} \right)^2 \tag{2.6}$$

to be the bootstrap estimates of $\boldsymbol{\beta}$ and $V(\hat{\boldsymbol{\beta}})$, respectively.

For many marginal models these estimates are consistent under the same conditions that admit consistency of $\hat{\boldsymbol{\beta}}$ and $\hat{V}_A(\hat{\boldsymbol{\beta}})$ (Burke and Gombay, 1991). Thus, $\hat{V}_A(\hat{\boldsymbol{\beta}})$ and $V^*_{nB}(\hat{\boldsymbol{\beta}})$ are often asymptotically equivalent. However, their performance may be quite different in small samples.

Notice that the resampling takes place from the *joint* distribution of $(t_1, \boldsymbol{x}_1)$ and $(t_2, \boldsymbol{x}_2)$, although modeling is done independently on the margins. This allows the resampling procedure to maintain the association between the two margins without explicit knowledge of the structure thereof. Furthermore, the relationship between t_m and $\boldsymbol{x}_m$ is maintained by this resampling plan. However, as was seen in the simulation study of Loughin (1993), the appropriateness of the proposed method depends critically on the assumption that the explanatory variables are random.

The problem of fixed explanatory variables is addressed in Section 3.

Often in real problems some of the event times are censored. Suppose censoring is random and follows a distribution $\mathcal{H}_{\boldsymbol{C}}(\cdot)$, where $\boldsymbol{C} = (C_1, C_2)$ are the censoring time random variables. Efron (1981a) demonstrated that when $M = 1$, resampling T^* and C^* from separate estimates of $\mathcal{F}_T$ and $\mathcal{H}_C$, respectively, and taking $T^{o*} = min(T^*, C^*)$ is equivalent to resampling pairs (T^{o*}, δ^*) directly from their two-dimensional distribution. However, for $M > 1$, obtaining estimates of $\mathcal{F}_T$ and $\mathcal{H}_C$ is not simple (see, e.g., Hanley and Parnes, 1983). On the other hand, resampling $\{(\boldsymbol{t}^{o(i)*}, \boldsymbol{x}^{(i)*}, \boldsymbol{\delta}^{(i)*})\}_{i=1}^n$ with replacement from the original data $\{(\boldsymbol{t}^{o(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{\delta}^{(i)*})\}_{i=1}^n$ is straightforward. Estimation of $\boldsymbol{\beta}$ and $V_A(\hat{\boldsymbol{\beta}})$ can proceed as before through (3), (5), and (6).

This method admits very general random censoring schemes, including type I (fixed endpoint) censoring. The distribution of the censoring indicators may depend arbitrarily on the distribution of $\boldsymbol{t}^o$ and $\boldsymbol{x}$. The special case of independence of $\boldsymbol{t}^o$ and $\boldsymbol{\delta}$ results in a particular type of dependence between $\mathcal{F}_T$ and $\mathcal{H}_C$. This will be discussed in greater detail in Section 3.

3. RESAMPLING WHEN EXPLANATORY VARIABLES ARE FIXED

Several authors have stressed the importance of proper handling of explanatory variables in resampling plans. Freedman (1981) and Hall (1992) draw the distinction between "regression" models with fixed explanatory variables and "correlation" models, where explanatory variables are random. When explanatory variables are fixed, as with treatment effects in comparative studies, they should not be jointly resampled with response values.

In linear regression problems with *iid* additive errors, for example, the ideal procedure involves resampling random errors and adding the resampled errors to the appropriate linear function of explanatory variables to obtain a new sample of responses. Since the errors are not directly observed, resampling must be done from some set of estimated residuals.

Freedman and Peters (1984) discuss resampling in a complex multivariate linear regression model. There they assume that *vectors* of errors, $(\epsilon_1, \ldots, \epsilon_M)$ are random variables drawn from an M-dimensional distribution. First, residuals are estimated within each margin, then resampling of error vectors is approximated by resampling from the resulting set of observed residual vectors.

We apply this resampling philosophy to the IWM for multivariate survival data

when explanatory variables are fixed. The first step is to define a set of residuals for this problem. In many cases this can be done using *generalized residuals*, which were developed by Cox and Snell (1968) for regression diagnostics.

Suppose there exists a set of functions $q^{(i)}$, $i = 1, \ldots, n$ such that

$$T^{(i)} = q^{(i)}(\boldsymbol{\beta}, \epsilon^{(i)}), \; i = 1, \ldots, n, \tag{3.1}$$

where $\epsilon^{(i)}$ $i = 1, \ldots, n$ are *iid* unobserved random variables. Suppose also that there exist functions $v^{(i)}$ such that

$$\epsilon^{(i)} = v^{(i)}(\boldsymbol{\beta}, T^{(i)}), \; i = 1, \ldots, n. \tag{3.2}$$

If $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$, then the generalized residual can be defined as

$$\hat{\epsilon}^{(i)} = v^{(i)}(\hat{\boldsymbol{\beta}}, T^{(i)}), \; i = 1, \ldots, n. \tag{3.3}$$

In linear regression this is achieved with

$$q^{(i)}(\boldsymbol{\beta}, \epsilon^{(i)}) = \boldsymbol{X}^{(i)}\boldsymbol{\beta} + \epsilon^{(i)}$$

and

$$\hat{\epsilon}^{(i)} = T^{(i)} - \boldsymbol{X}^{(i)}\hat{\boldsymbol{\beta}}.$$

In the multivariate survival problem many marginal models can be expressed in forms like (1) and its dual, (2). Cox and Snell explicitly consider examples for the exponential and Weibull distributions.

When a parametric form is adopted as a marginal survival model, an alternative method of residual construction is the application of the probability integral transformation to provide uniform residuals. When no censoring is present,

$$\hat{\epsilon}_m^{(i)} = 1 - S_m(t_m^{(i)}; \boldsymbol{x}_m^{(i)}, \hat{\boldsymbol{\beta}}_m), \; i = 1, \ldots, n \tag{3.4}$$

yields a set of exchangeable observations which are approximately uniformly distributed on $(0,1)$ for each $m = 1, \ldots, M$. Vectors of errors are then considered to be *iid* from a multivariate distribution with uniform margins. Multivariate uniform distributions are often used as bases for creating very general multivariate distributions (see, e.g., Barnett, 1980), and so this approach admits many models.

Once vectors of residuals have been created for all subjects in the study, resampling and data reconstruction are straightforward. Vectors $\{\boldsymbol{\epsilon}^{(i)*}\}_{i=1}^{n}$ are selected with replacement from $\{\hat{\boldsymbol{\epsilon}}^{(i)}\}_{i=1}^{n}$, and a new set of failure times $\{\boldsymbol{T}^{(i)*}\}_{i=1}^{n}$ are generated by inserting $\epsilon_m^{(i)*}$ and $\hat{\boldsymbol{\beta}}_m$ into (1) for each margin. These are then used to create bootstrap estimates $\boldsymbol{\beta}_m^*$ as in Section 2.

Censored data is handled in much the same way as in Section 2. A time which is censored gives rise to a residual which itself is a censored observation from its distribution. Hence, vectors of estimated residuals are associated with vectors of censoring indicators, and these are resampled jointly. Loughin (1993) shows that resampling censoring indicators with residuals in this manner implies that the distribution of censoring times must be a function of the conditional distribution of failure times, given the explanatory variables. When this requirement is not satisfied, Loughin (1993) suggests that this resampling plan is probably still adequate when censoring is light or when the amount of censoring is similar for all values of the explanatory variables.

When some or all of the explanatory variables correspond to group memberships and the distribution of censoring is believed to differ among groups, a cruder version of this resampling plan is possible. Residuals within groups might then be considered to be from different populations, possibly subject to different censoring mechanisms.

In this case, the estimated residuals should be resampled along with their censoring indicators separately within groups. If the group membership is the only explanatory variable, this amounts to resampling the times and censoring indicators jointly. This approach was used by Efron and Gong (1983) and by Karrison (1990) in analyses of univariate failure time data.

A special model that has gained great favor since its introduction by Cox (1972) is the proportional hazards model. Development of residuals for this model deserves special attention here. Suppose we can write the marginal hazard function as

$$h_m(t_m; \boldsymbol{x}_m) = h_{0m}(t_m) g_m(\boldsymbol{x}_m, \boldsymbol{\beta}_m), \tag{3.5}$$

where $g_m(\boldsymbol{x}_m, \boldsymbol{\beta}_m)$ is a positive *relative risk function* and $h_{0m}(t_m)$ is the *baseline hazard function* for a subject with $g_m(\boldsymbol{x}_m, \boldsymbol{\beta}_m) = 1$. Estimation of $\boldsymbol{\beta}_m$ is possible without specification of h_{0m} through the *partial likelihood*

$$L_{PH}(\boldsymbol{\beta}_m) = \prod_{i=1}^{k} \frac{g_m(\boldsymbol{x}_m^{(i)}, \boldsymbol{\beta}_m)}{\sum_{l \in \mathcal{R}^{(i)}} g_m(\boldsymbol{x}_m^{(l)}, \boldsymbol{\beta}_m)}. \tag{3.6}$$

where the product is taken over the k distinct failure times in the data, and $\mathcal{R}^{(i)}$ is the set of all subjects at risk at the time of the i^{th} ordered failure.

Kalbfleisch and Prentice (1973) noted that this likelihood remains invariant under the group of monotone increasing transformations of t when g_m is independent of time and h_{0m} is strictly positive over all open intervals. Loughin used this fact to create generalized residuals for use in bootstrapping the estimator resulting from maximizing (6). The scale-invariance of this likelihood allows reconstruction of "times" in a different scale from that of the original data, thus avoiding the need to specify

a form for h_{0m}. This method will be used in the example in Section 4 and in the simulations of Sections 5 and 6.

4. APPLICATION TO VIRAL POSITIVITY DATA

To illustrate the proposed method and provide a basis for a comparative Monte Carlo study, the viral positivity data of Wei, Lin, and Weissfeld are analyzed. These data come from a clinical trial designed to test the effectiveness of two dosage levels of the drug ribavirin as a retroviral treatment for patients with AIDS. Thirty-six patients were randomly assigned either a placebo or a low or high dosage of the active drug. Hence, the two dummy variables corresponding to treatment group assignment are fixed by design, rather than sampled from some population. Blood serum samples were collected at three monthly follow-up visits. These samples were observed until a viral marker surpassed a tolerance level. The time (in days) at which this event occurred was recorded for each sample.

. The structure of the study provides the opportunity to observe three distinct event times on each patient. However, the data are subject to random censoring due to a variety of causes, including contamination of the sample and failure of the patient to provide a sample. It may be reasonable to assume that these causes affect all three groups equally. In that case, a residual resampling plan from Section 3 can be applied.

Wei, Lin, and Weissfeld fit a proportional hazards model in each margin. They

Table 4.1: Parameter estimates for viral positivity data

Month	Dose	IWM Estimate	Bootstrap β^*	Bootstrap BIAS*	Bootstrap $\beta^*_{BC}{}^a$
1	LOW	-1.39	-1.53	-0.14	-1.25
1	HIGH	-0.93	-0.95	-0.01	-0.91
2	LOW	-0.66	-0.82	-0.16	-0.50
2	HIGH	0.02	0.06	0.04	-0.02
3	LOW	-0.62	-0.72	-0.11	-0.52
3	HIGH	-0.33	-0.39	-0.06	-0.27

[a]Bias-corrected.

assume a loglinear form for the relative risk,

$$g(\boldsymbol{x}, \boldsymbol{\beta}) = e^{\boldsymbol{x}\boldsymbol{\beta}},$$

and maximize (6) to estimate parameter values. As an estimate of $V(\hat{\boldsymbol{\beta}})$ they use a matrix of the form (4), with modifications to the individual scores to account for the fact that they are not *iid*. We repeat this analysis with the exception that a bootstrap is used to obtain estimates of the bias and variance of $\hat{\boldsymbol{\beta}}$. One thousand bootstrap replicates were drawn by the methods of Section 3 using the proportional hazards residuals proposed by Loughin (1993). Results from the two estimation methods are given in Tables 4.1–4.3.

Bootstrap estimates of bias in Table 4.1 suggest that all of the original parameter estimates for the low dose are biased away from zero. Bias correction reduces the magnitude of these values. Table 4.2 shows that the bootstrap provides larger estimates of variance for $\hat{\beta}$ than does the robust estimator, but correlations provided by these two estimators are not substantially different. Ninety-percent confidence limits

Table 4.2: Variance estimates for viral positivity data

ROBUST COVARIANCE MATRIX

Month	Param	Covariance Estimates					
1	LOW	.255					
1	HIGH	.075	.136				
2	LOW	.051	.026	.287			
2	HIGH	.017	.041	.119	.167		
3	LOW	.107	.046	.133	.076	.257	
3	HIGH	.061	.091	.091	.069	.114	.229

BOOTSTRAP COVARIANCE MATRIX

Month	Param	Covariance Estimates					
1	LOW	.331					
1	HIGH	.148	.227				
2	LOW	.078	.043	.342			
2	HIGH	.044	.071	.127	.259		
3	LOW	.140	.040	.076	.035	.365	
3	HIGH	.080	.101	.040	.081	.184	.328

for the individual components of β appear in Table 4.3. Both the normal-based intervals with robust variance estimates and the bias-corrected percentile intervals of Efron (1981b) provide essentially the same inferences. Both levels of ribavirin appear to increase time until viral positivity is attained in the first month but not in later months.

The simulation results in Section 6 show that the relationships noted here between the bootstrap and robust estimates of $V(\hat{\beta})$ are consistent with the general patterns observed for these methods. The bootstrap provides a correction for the

Table 4.3: 90% confidence intervals for viral positivity data

Month	Param	Method	Confidence Limits Lower	Upper
1	LOW	Norm-Rob [a]	-2.22	-0.55
		PCT-BC [b]	-2.23	-0.46
	HIGH	Norm-Rob	-1.54	-0.32
		PCT-BC	-1.80	-0.25
2	LOW	Norm-Rob	-1.54	0.22
		PCT-BC	-1.46	0.37
	HIGH	Norm-Rob	-0.65	0.69
		PCT-BC	-0.88	0.78
3	LOW	Norm-Rob	-1.45	0.21
		PCT-BC	-1.60	0.32
	HIGH	Norm-Rob	-1.12	0.46
		PCT-BC	-1.19	0.66

[a] Normal-based intervals with robust variance estimates.

[b] Bias-corrected percentile intervals of Efron (1981b).

tendency, noted by Johnson *et al.* (1982) and Loughin (1993), for the usual parameter estimate $\hat{\beta}$ to be biased away from zero. Use of the bootstrap also appears to provide improved variance estimates.

5. DESCRIPTION OF SIMULATION STUDY

A Monte Carlo simulation study was designed to examine under a variety of conditions the properties of several variance estimators for the multivariate survival problem. Motivated by the previous example, the context is that of univariate marginal proportional hazards models in each of three dimensions with explanatory variables which are fixed by design. The residual resampling plan of Section 3 is used to obtain a bootstrap estimate of $V(\hat{\beta})$, and two asymptotic variance estimates are calculated: the robust matrix of Wei, Lin, and Weissfeld and the negative inverse of the block diagonal matrix containing the observed information matrices for the three margins. The last of these does not consider cross-marginal covariances.

Multivariate data with specified rank correlations, τ, and proportional hazards margins were generated for the analysis through the following steps (see Johnson (1987) for technical details). First uniform pseudo-random numbers were generated using the algorithm of Wichman and Hill (1982). These were transformed into independent standard normals using the Box-Muller approach. Values of the correlation coefficients, ρ, for the normal distribution were found from the specified rank correlations through

$$\rho = sin\frac{\pi}{2}\tau$$

as given by Kendall (1955). Multivariate normal data were then created using the

Choleski factorization on the correlation matrix. These were then transformed into a multivariate uniform distribution possessing the desired rank correlations by applying the standard normal cdf in each margin. Data from the margins of this distribution were converted into *probability-scale* "times" from a proportional hazards distribution through the relation given by Loughin (1993),

$$T_m = 1 - [U_m]^{(1/g_m(\boldsymbol{x}_m,\boldsymbol{\beta}_m))}, \; m = 1, \ldots, M,$$

where U_m represents the uniform variate from margin m.

Monte Carlo runs were made under a variety of conditions. Samples of size 24, 50, and 100 were generated. These were subject to random censoring from the distribution of Koziol and Green (1976). This model for the censoring is compatible with the residual resampling plans and is quite easy to implement in the generation of proportional hazards data. Amounts of censoring depended on sample size. For $n = 24$ only 0% or 20% censoring was used; with $n = 50$ an additional run with 50% censoring was done; and at $n = 100$, 80% censoring was also included, simulating a preliminary analysis of an unfinished study. Data were generated with $\tau = 0$, the independence case, and with $\tau = .6$, representing a moderately high association value. Note that this represents underlying association of the *residual* distribution, not the distribution of times. The latter is influenced by the effects of explanatory variables in different margins and may have somewhat different levels of association.

The previous simulation studies mentioned in Section 2 both noted deterioration in the performance of asymptotic estimators with increasing numbers of explanatory variables. Hence, two separate sets of runs were made. The first, using all nine combinations of sample sizes and censoring fractions, included only one explanatory variable, a 0–1 "treatment" indicator. All samples were split so that half of the

sample was assigned each value. Thus, the explanatory variable is the same in all margins. Parameter values for the three margins were chosen to be 1.0, .50, and .25. In the second set of runs a five-group problem was created, as in a study where subjects are assigned to different concentrations or dosages of a treatment. One-fifth of each sample was assigned membership in each group. Hence, there were four parameters to be estimated in each margin. The parameters were assigned arbitrary values between 0 and 1. Due to the increased potential for numerical instability (see below) only samples of size 50 with 0% or 20% censoring and size 100 with 0%, 20%, or 50% censoring were generated.

As with any Monte Carlo study of the proportional hazards model, the danger exists that samples with *monotone likelihood* will be generated. This occurs when any of the explanatory variables, or certain functions thereof, are monotone with respect to the ordered failure times; for example, in the two-group case, when the last failure in one group precedes the first failure in the other group. There is then no basis on which to estimate the proportionality of the hazards, and so $\hat{\beta}$ goes to $\pm\infty$. Bryson and Johnson (1981) recommend replacing such cases with new samples and making estimates of β and $V(\hat{\beta})$ conditional on finite $\hat{\beta}$. Loughin (1993) suggests that a high proportion of replicates with monotone likelihood in a bootstrap analysis is an indication that the proportional hazards model may not be appropriate, since an appreciable mass of the sampling distribution of $\hat{\beta}$ is placed at $\pm\infty$. A variance for $\hat{\beta}$ estimated by any method therefore has little real meaning under these circumstances. In the present study, the recommendation of Bryson and Johnson was adopted: Monte Carlo trials and bootstrap replicates were replaced if a monotone likelihood was detected.

Because this approach can cause difficulties in interpreting simulation results information on the incidence of monotone likelihood in this study is provided in Table 5.1. For each sample size considered for the problem of one parameter per margin, too much censoring produces significant biases in estimates of β and $V(\hat{\beta})$ due to high prevalence of monotone likelihood in the bootstrap samples. Simulation results for cases where more than 50% of the bootstrap samples exhibited monotone likelihood are therefore not included in the following discussion.

Table 5.1: Monotone likelihood cases in simulation study

Number of Params	n	Censoring	τ	M.C. Trials Replaced	Bootstrap Replicates Replaced [a] Mean	Median	Max	%=0 [b]
1	24	0%	0	0	14	1	844	37.1
			.6	0	10	1	857	48.3
		20%	0	3	44	11	1254	5.0
			.6	4	37	8	1134	10.9
	50	0%	0	0	0	0	4	99.0
			.6	0	0	0	9	99.5
		20%	0	0	0	0	23	94.5
			.6	0	0	0	18	94.4
		50%	0	1	10	2	536	26.9
			.6	0	12	2	1193	29.6
	100	0%	0	0	0	0	0	100.0
			.6	0	0	0	0	100.0
		20%	0	0	0	0	0	100.0
			.6	0	0	0	0	100.0
		50%	0	0	0	0	7	98.5
			.6	0	0	0	7	98.9
		80%	0	10	48	21	1071	2.3
			.6	8	47	17	1537	3.6
4	50	0%	0	0	0	0	4	99.0
			.6	0	0	0	9	99.5
		20%	0	0	0	0	23	94.5
			.6	0	0	0	18	94.4
	100	0%	0	0	0	0	0	100.0
			.6	0	0	0	0	100.0
		20%	0	0	0	0	0	100.0
			.6	0	0	0	0	100.0
		50%	0	0	0	0	7	98.5
			.6	0	0	0	7	98.9

[a] Numbers tabulated over 1000 Monte Carlo trials.

[b] Percentage of trials with no bootstrap replicates replaced.

6. SIMULATION RESULTS

The simulation study by Loughin (1993) showed that the residual bootstrap for univariate proportional hazards models provides substantial reduction in mean squared error (MSE) for estimating β when compared to the standard partial likelihood method. Since the models in the present study are fit to the margins independently, similar results are observed, and no further discussion regarding the estimation of β is presented here.

Results from the estimation of $V(\hat{\beta})$ for the case with one-parameter marginal models are presented in Table 6.1. Variance and covariance estimates are averaged across both the three marginal or cross-marginal estimates, respectively, and the 1000 Monte Carlo samples. Results for the three margins are sufficiently similar to permit this summarization; only a slight systematic increase in variance was noted as parameter values increased. The finite sample variances, which represent the estimated true variance $V(\hat{\beta})$, are the observed variances of the estimates of β from the 1000 Monte Carlo trials. Mean squared errors for the three variance estimation methods are calculated with respect to the finite sample variances. Both of these quantities are also averaged across margins.

The estimates of within-marginal variance obtained by each method are virtually identical for the two levels of association. This is to be expected, since the corre-

sponding IWM marginal models are the same. In each of the six censoring-by-sample size combinations, the bootstrap provides the largest estimates of variance, while the robust estimates are the smallest. In most cases the bootstrap estimates are nearer the finite sample variances than are the robust estimates. Mean squared errors for the three estimators are comparable, although inverting the IWM information matrix tends to provide the smallest MSEs.

In the estimation of the cross-marginal covariances, differences among the methods are more evident. When $\tau = 0$, the IWM, which correctly assigns the value 0 to the covariances, has uniformly smaller MSE than the other methods. When there is association among the margins, however, the IWM method is obviously inadequate. Both the bootstrap and the robust methods capture the effect of the association on the covariances of parameters in different margins. Bootstrap estimates of these covariances are nearly unbiased. Although the robust estimates of covariance tend to be too small, the correlation structure is nonetheless quite close to that of the finite sample variance matrix. The MSEs of these two methods do not differ greatly.

Table 6.2 shows the results from the estimation of $V(\hat{\beta})$ in the case of four-parameter marginal models. The 78 elements of the variance matrix are classified into four groups: variances, within-margin covariances, across-margin covariances of the same treatment parameter, and across-margin covariances of different treatment parameters. As in the previous case, entries in Table 6.2 represent averages taken over all elements within a classification.

Estimates of both variance and within-margin covariance are again essentially the same for the two association levels. However, in this problem there is a distinct advantage to the bootstrap estimation. in each sample size and censoring fraction,

bootstrap variance estimates are nearer to the finite sample variance than either of the asymptotic estimates. The bootstrap estimates tend to be slightly conservative, while the others are generally biased low. The corresponding MSEs for the bootstrap variances are uniformly lower than those for the other methods. Mean squared errors for the bootstrap estimates range from 13–65% smaller than the corresponding MSEs for the inverse information estimates, and they are as much as 85% smaller than those for the robust estimator. Similar but smaller differences are present for within-margin covariances in most of the combinations of censoring and sample size. Here, too, the bootstrap and IWM estimates have smaller MSE that the robust estimates.

In the estimation of cross-marginal covariances, the bootstrap still compares favorably to the robust method. In the case of no association, bootstrap estimation offers a 40–50% reduction in MSE over the robust method for estimating covariances for the same treatment parameter. Further reduction, to nearly 70%, occurs in estimating the cross-marginal covariances corresponding to different parameters. When there is association among the margins, bootstrap estimates of both cross-marginal covariances are noticeably closer to the finite sample values for most combinations of sample size and censoring fraction. Improved MSEs are evident in each situation, ranging from 55–80% smaller than those from the robust method, with greater reductions in the less heavily censored samples.

Confidence intervals for the parameters are formed using the bias-corrected percentile method of Efron (1981b) for bootstraps and the standard normal approximation for the inverse IWM information and robust variance matrices. Loughin (1993) compared the 90% bootstrap and inverse information intervals and found that the latter are too short in the univariate four-parameter estimation problem, resulting

in coverage below the nominal level. The bias-corrected percentile intervals provide coverage that is slightly above the nominal level. Similar results are in evidence in the present study. Additionally, intervals formed using the robust variance estimates are shorter than those from the inverse IWM information, and their coverage is generally a little lower, though never by more than 2%.

Table 6.1: Simulation results: Sampling variance estimates with mean squared errors for estimation in one-parameter margins

				Within Margins		Across Margins	
τ	n	Censor	Method	VAR	MSE[a]	COV	MSE
0	24	0	FSV[b]	.255	—	.002	—
			IWM[c]	.215	532	0	1
			ROB[d]	.198	692	.002	194
			BOOT[e]	.253	254	.001	201
	50	0	FSV	.101	—	.001	—
			IWM	.095	21	0	1
			ROB	.090	32	-.000	18
			BOOT	.102	31	-.000	18
		20	FSV	.118	—	.000	—
			IWM	.118	29	0	1
			ROB	.112	32	-.000	29
			BOOT	.128	68	-.000	29
	100	0	FSV	.048	—	-.001	—
			IWM	.046	1	0	0
			ROB	.044	3	-.000	2
			BOOT	.047	3	-.000	2
		20	FSV	.059	—	-.001	—
			IWM	.057	3	0	1
			ROB	.056	5	.001	4
			BOOT	.059	5	.000	4
		50	FSV	.097	—	.003	—
			IWM	.093	18	0	3
			ROB	.090	24	.000	11
			BOOT	.097	29	.000	11

[a] All MSEs are $\times 10^{-5}$.

[b] Finite Sample Variance of the 1000 Monte Carlo estimates of β.

[c] Inverse of (-) the IWM information matrix.

[d] Robust matrix of Wei, Lin, and Weissfeld (1989).

[e] Multivariate residual bootstrap.

Table 6.1 (Continued)

				Within Margins		Across Margins	
τ	n	Censor	Method	VAR	MSE	COV	MSE
.6	24	0	FSV	.240	—	.168	—
			IWM	.212	358	0	2837
			ROB	.195	482	.133	225
			BOOT	.248	234	.158	115
	50	0	FSV	.097	—	.070	—
			IWM	.094	17	0	486
			ROB	.090	24	.064	14
			BOOT	.102	31	.070	11
		20	FSV	.123	—	.076	—
			IWM	.119	34	0	571
			ROB	.113	43	.070	21
			BOOT	.129	61	.076	17
	100	0	FSV	.046	—	.033	—
			IWM	.046	1	0	111
			ROB	.044	2	.032	2
			BOOT	.047	3	.034	1
		20	FSV	.058	—	.036	—
			IWM	.057	3	0	132
			ROB	.056	4	.035	2
			BOOT	.059	5	.037	2
		50	FSV	.094	—	.049	—
			IWM	.093	18	0	236
			ROB	.090	21	.047	6
			BOOT	.097	28	.049	7

Table 6.2: Simulation results: Sampling variance estimates with mean squared errors for estimation in four-parameter margins

				Within Margins				Across Margins			
				Variance		Covariance		Same x		Different x	
τ	n	Cens	Method	VAR	MSE[a]	COV	MSE	COV	MSE	COV	MSE
0	50	0	FSV[b]	.278	—	.147	—	-.003	—	-.002	—
			IWM[c]	.226	340	.116	145	0	8	0	7
			ROB[d]	.219	697	.115	518	.001	267	-.000	163
			BOOT[e]	.285	120	.150	97	.002	161	.001	55
		20	FSV	.342	—	.170	—	.003	—	.000	—
			IWM	.291	520	.150	195	0	11	0	7
			ROB	.280	1087	.146	545	.000	416	-.000	250
			BOOT	.364	386	.192	269	.001	232	.001	79
	100	0	FSV	.118	—	.060	—	.001	—	.002	—
			IWM	.108	16	.055	7	0	2	0	2
			ROB	.106	61	.054	31	-.000	30	-.000	17
			BOOT	.121	10	.063	8	-.000	17	-.000	7
		20	FSV	.155	—	.084	—	.001	—	.001	—
			IWM	.137	58	.070	33	0	3	0	2
			ROB	.133	125	.068	69	-.000	47	-.000	27
			BOOT	.151	26	.079	17	-.000	25	-.000	10
		50	FSV	.242	—	.124	—	-.004	—	-.001	—
			IWM	.228	246	.116	134	0	5	0	2
			ROB	.221	369	.114	214	.000	132	.000	73
			BOOT	.257	214	.133	103	.000	70	-.000	24

[a] All MSEs are x10^{-5}.

[b] Finite Sample Variance of the 1000 Monte Carlo estimates of β.

[c] Inverse of (-) the IWM information matrix.

[d] Robust matrix of Wei, Lin, and Weissfeld (1989).

[e] Multivariate residual bootstrap.

Table 6.2 (Continued)

				Within Margins				Across Margins			
				Variance		Covariance		Same x		Different x	
τ	n	Cens	Method	VAR	MSE	COV	MSE	COV	MSE	COV	MSE
.6	50	0	FSV	.273	—	.143	—	.199	—	.103	—
			IWM	.226	272	.116	113	0	3973	0	1069
			ROB	.219	612	.117	314	.160	435	.085	230
			BOOT	.284	119	.151	96	.198	85	.105	47
		20	FSV	.345	—	.176	—	.218	—	.111	—
			IWM	.291	584	.150	255	0	4759	0	1235
			ROB	.279	1159	.145	602	.175	578	.091	296
			BOOT	.365	410	.193	264	.217	158	.114	82
	100	0	FSV	.123	—	.066	—	.093	—	.050	—
			IWM	.108	29	.055	16	0	870	0	251
			ROB	.106	71	.055	39	.079	56	.041	30
			BOOT	.121	10	.063	8	.088	12	.046	7
		20	FSV	.150	—	.078	—	.096	—	.050	—
			IWM	.137	41	.070	20	0	940	0	253
			ROB	.134	100	.069	54	.086	57	.045	31
			BOOT	.152	26	.079	15	.095	17	.049	8
		50	FSV	.242	—	.123	—	.122	—	.061	—
			IWM	.228	229	.116	122	0	1498	0	375
			ROB	.221	355	.113	202	.114	112	.059	73
			BOOT	.257	196	.133	94	.128	49	.066	25

7. DISCUSSION

The results in Section 6 provide encouragement for the use of the bootstrap in multivariate survival problems. Both the bootstrap and robust estimation methods give very good approximations to the true variance matrix $V(\hat{\beta})$ for the problem studied in the simulations. However, the magnitude of the improvement in mean squared errors resulting from bootstrap estimation of $V(\hat{\beta})$ suggests that the bootstrap estimates are more reliable. The bootstrap also offers estimates of the bias of $\hat{\beta}$. For proportional hazards models, Loughin (1993) shows that bootstrap bias-corrected estimates are more efficient in small samples than uncorrected partial likelihood estimates, providing a reduction in MSE beyond the effect of the bias. One advantage to model-robust matrices like $\hat{V}_A(\hat{\beta})$ is their consistency even when the marginal models are misspecified (Huber, 1967; Royall, 1986; Lin and Wei, 1989). It is not clear how reliable bootstrap methods are in such circumstances.

While the features of IWM approach are compelling, as noted in Section 2 there can be expected to be a loss of efficiency compared to a full parametric approach. Huster *et al.* study the asymptotic relative efficiency of their IWM estimator when the times follow the bivariate distribution of Clayton (1978) and Oakes (1982). They find that the efficiency of their estimator is reasonably high when correlation is low, but that this efficiency decreases to below 50% when association becomes high. At

a correlation of .64, the asymptotic relative efficiency ranges between 70–90% under varying conditions of practical interest. Further loss of efficiency may occur when semiparametric marginal models like (5) are used. Efron (1977) studied the proportional hazards regression estimator $\hat{\beta}$ and found that it is fully asymptotically efficient for $\beta = 0$, with decreasing efficiency as β moves away from zero (see Kalbfleisch and Prentice, 1980, for more details). Still, the difficulties associated with fitting fully-specified multivariate models may make a slight loss of efficiency a small price to pay for the flexibility and intuitive appeal of the IWM approach.

As with many other multivariate problems, the amount of data required for reliable estimation increases at a rate faster than linearity with the number of dimensions. This "curse of dimensionality" may also affect bootstrap methods in which many explanatory variables are resampled along with responses. For example, in the univariate simulation study of Loughin (1993), the performance of the bootstrap method which resamples vectors of explanatory variables along with failure times deteriorates badly with increasing numbers of fixed explanatory variables. Further investigation of this aspect of bootstrap performance is needed. It is noted that the method of resampling residuals avoids this difficulty and should be used whenever the explanatory variables are nonrandom.

Finally, extensions of the residual resampling methods of Section 3 must still be developed for a wider variety of sampling schemes. In many cases, resampling censoring indicators along with residuals implies a certain structure for the censoring time distribution. For the univariate proportional hazards model studied by Loughin (1993), for example, the residual bootstrap did not perform well in a simulation where the data were subject to Type I censoring. Work is underway to address this problem

in the univariate setting; at present it is unclear how to handle fixed censoring in the multivariate setting.

BIBLIOGRAPHY

Barnett, V. (1980). Some Bivariate Uniform Distributions. *Communications in Statistics: Theory and Methods*, **9**, 453–461.

Bryson, M. C. and Johnson, M. E. (1981). The Incidence of Monotone Likelihood in the Cox Model. *Technometrics*, **23**, 381–383.

Burke, M. D. and Gombay, E. (1991). The Bootstrapped Maximum Likelihood Estimator with an Application. *Statistics and Probability Letters*, **12**, 421–427.

Cox, D. R. (1972). Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–202.

Cox, D. R. and Snell, E. J. (1968). A General Definition of Residuals (with discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 248–265.

Crowder, M. (1989). A Multivariate Distribution with Weibull Connections. *Journal of the Royal Statistical Society, Series B*, **58**, 93–107.

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, **72**, 557–565.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, **7**, 1–26.

Efron, B. (1981a). Censored Data and the Bootstrap. *Journal of the American Statistical Association*, **76**, 312–319.

Efron, B. (1981b). Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics*, **9**, 139–172.

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, Jackknife, and Crossvalidation. *American Statistician*, **37**, 36–48.

Freedman, D. A. (1981). Bootstrap Regression Models. *Annals of Statistics*, **9**, 1218–1228.

Freedman, D. A. and Peters, S. C. (1984). Bootstrapping a Regression Equation: Some Empirical Results. *Journal of the American Statistical Association*, **79**, 97–106.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer-Verlag.

Hanley, J. A. and Parnes, M. N. (1983). Nonparametric Estimation of A Multivariate Distribution in the Presence of Censoring. *Biometrics*, **39**, 129–139.

Hougaard, P. (1987). Modeling multivariate Survival. *Scandinavian Journal of Statistics*, **14**, 291–304.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimated under Nonstandard Conditions. *Proceedings of the Fifth Berkeley Symposium*, 221-233.

Huster, W. J., Brookmeyer, R., and Self, S. G., (1989). Modeling Paired Survival Data with Covariates. *Biometrics*, **45**, 145–156.

Johnson, M. E. (1987). *Multivariate Statistical Simulation.* New York: John Wiley and Sons.

Johnson, M. E., Tolley, H. D., Bryson, M. C., and Goldman, A. S. (1982). Covariate Analysis of Survival Data: A Small Sample Study of Cox's Model. *Biometrics*, **38** 685–698.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal Likelihoods Based of Cox's Regression and Life Model. *Biometrika*, **60**, 267–278.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: John Wiley and Sons.

Karrison, T. (1990). Bootstrapping Censored Data with Covariates. *Journal of Statistical Computation and Simulation*, **36**, 195–207.

Kendall, M. G. (1955). *Rank Correlation Methods*, 2nd ed. London: Charles Griffin and Company.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises Statistic for Randomly Censored Data. *Biometrika*, **63**, 465–474.

Lin, D. Y. and Wei, L. J. (1989). The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*, **84**, 1074–1078.

Loughin, T. M. (1993). Bootstrap Applications in Proportional Hazards Models. *Ph.D. Dissertation, Iowa State University*

Royall, R. M. (1986). Model Robust Confidence Intervals using Maximum Likelihood Estimators. *International Statistical Review*, **54**, 221–226.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of*

the American Statistical Association, **84**, 1064–1073.

Wichman, B. A. and Hill, I. D. (1982). An efficient and portable pseudo-random number generator. *Applied Statistics*, **31**, 188–190.

PAPER III.

A SEMIPARAMETRIC BOOTSTRAP FOR PROPORTIONAL HAZARDS REGRESSION MODELS

ABSTRACT

A bootstrap resampling plan is developed for the proportional hazards estimator of Cox (1972) for the case when explanatory variables are not random. Instead of resampling observed times, the proposed plan resamples from the Uniform(0,1) distribution of probability integral transformations of conditional failure times. Since the partial likelihood is invariant to monotone increasing transformations of the failure times, the analysis may be performed without transforming resampled values back into the time scale in which the original data are measured. The resampling method is easily adapted to a wide variety of censoring schemes. Applications to the childhood leukemia data of Cox (1972) and the carcinogenesis data of Kalbfleisch and Prentice (1980) are given. A simulation study provides comparisons with the standard partial likelihood estimation procedures.

1. INTRODUCTION

Consider a study in which the time is measured until an event occurs (such as death or failure of a study subject). Let T be a random variable corresponding to time to failure. Let $\boldsymbol{x}$ correspond to a set of explanatory variables. It is assumed throughout that $\boldsymbol{x}$ represents a set of nonrandom values fixed by the experimental design, such as treatment assignments in a designed clinical trial. Suppose we wish to assess the effects of different values of $\boldsymbol{x}$ on the distribution of T. A convenient quantity to model in such studies is the *hazard function*,

$$h(t) = \lim_{\Delta \to 0} \frac{Pr(t < T < t + \Delta | T > t)}{\Delta}.$$

Cox (1972) proposed the *proportional hazards* factorization

$$h(t|\boldsymbol{x}) = h_0(t) g(\boldsymbol{x}, \boldsymbol{\beta}), \tag{1.1}$$

where $g(\boldsymbol{x}, \boldsymbol{\beta})$ is a positive *relative risk* function of the explanatory variables and some unknown parameters $\boldsymbol{\beta}$, and $h_0(t)$ is the *baseline hazard function* for a subject with unit relative risk. Given an appropriate form for $g(\boldsymbol{x}, \boldsymbol{\beta})$, estimation depends only on the assumptions made about the form of the baseline hazard function.

If one wishes to avoid parametric assumptions about $h_0(t)$, then the estimation of $\boldsymbol{\beta}$ can be based on a partial likelihood function derived by Cox (1972, 1975). Suppose k distinct failure times are observed in a study on n individuals. Denote the

ordered failure times by $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$, and define $\mathcal{R}_i$, the *risk set* at the i^{th} ordered failure time, as the set of indices $\{j : t_j \geq t_{(i)}; j = 1, \ldots, n\}$, where t_j denotes the failure time of the j^{th} individual. Then, conditional on the individuals failing prior to time $t_{(i)}$ (or, equivalently, on $\mathcal{R}_i$), and on the fact that a single failure occurred *at* $t_{(i)}$, the probability that the i^{th} ordered failure is provided by individual j is simply $\frac{h(t_{(i)}|\boldsymbol{x}_j)}{\sum_{l \in \mathcal{R}_i} h(t_{(i)}|\boldsymbol{x}_l)}$. Application of (1) with the observed ordering of failure times yields the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{g(\boldsymbol{x}_i, \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}_i} g(\boldsymbol{x}_l, \boldsymbol{\beta})}. \tag{1.2}$$

Cox suggested the relative risk function $g(\boldsymbol{x}, \boldsymbol{\beta}) = exp(\boldsymbol{x}\boldsymbol{\beta})$, which has been adopted in the bulk of the literature.

Simulation studies of the Cox partial likelihood estimator $\hat{\boldsymbol{\beta}}$ by Johnson *et al.* (1982) and Loughin (1993) have shown that it is biased away from zero in small samples. Furthermore, the usual asymptotic estimate of $V(\hat{\boldsymbol{\beta}})$ exhibits a downward bias, which becomes more pronounced as the dimension of $\boldsymbol{\beta}$ increases. Different bootstrap methods have been used to estimate $\boldsymbol{\beta}$ and to obtain corresponding standard errors and confidence intervals. Those by Efron and Gong (1983) and Loughin (1993) are intended for problems with fixed explanatory variables, while the method used by Chen and George (1985), Efron and Tibshirani (1986), and Altman and Andersen (1989) requires that the explanatory variables are random. The methods of Efron and Gong and of Loughin differ in that the former assumes that the explanatory variables represent membership in groups of sufficient size that times can be resampled separately within each group. The latter method resamples from a set of estimated residuals, as is done in analogous linear regression problems (see Efron,

1979; Freedman, 1981; and Wu, 1986). The method employed by the other authors involves resampling times and explanatory variables jointly, and thus is appropriate only when explanatory variables are random. All of these resampling schemes induce tied failure times in the bootstrap samples. The usual likelihood (2) is not adequate when ties are present, since the numerator in each factor corresponds to the relative risk of a *single individual's* failure at each ordered failure time. When ties are due to some slight grouping of what could otherwise be thought of as distinct times, the true likelihood can be found. However, this can quickly become computationally cumbersome as the number of ties increases. Alternatively, an approximation to the likelihood proposed by Breslow (1974) has become widely adopted because of its computational ease. However, Farewell and Prentice (1980) find that the resulting estimator of β can be badly biased when there are large numbers of ties.

In Section 2 of this paper, a bootstrap method is proposed which allows resampling from a *continuous* distribution, thus avoiding the potential for tie-induced biases in the resulting bootstrap estimators. If there is no censoring, then the resampling plan for the estimation of β is equivalent to randomly selecting new sets of observations from a proportional hazards distribution with parameter $\hat{\beta}$. Two examples of the use of this method, involving different censoring mechanisms, are provided in Section 3. A series of simulations in Section 4 show that this new bootstrap estimator reduces bias in the estimation of β and $V(\hat{\beta})$, and provides accurate coverage for regression parameter confidence intervals. Possible modifications and extensions of this new method are discussed in Section 5.

2. A SEMIPARAMETRIC BOOTSTRAP

A key feature of Cox's likelihood (2) is that it allows for estimation of regression parameters without specification of the form of the baseline hazard function, $h_0(t)$. Thus, assessments of treatment effects, for example, can be obtained without potentially restrictive distributional assumptions. Bootstrap methods can be used to reduce the bias of the resulting estimators and obtain more accurate standard errors and confidence intervals, but censoring mechanisms must be carefully considered.

Bootstrap estimation of the regression parameters β in the proportional hazards model requires a sample-based approximation to $\mathcal{G}_n(\hat{\beta}|F_\beta)$, the sampling distribution of the estimator $\hat{\beta}$ that maximizes (2) for random samples of size n from the true unknown proportional hazards distribution F_β. Let $\tilde{F}_n$ be some estimate of F_β based on the observed data. The bootstrap approximation to $\mathcal{G}_n(\hat{\beta}|F_\beta)$ is simply $\mathcal{G}_n(\beta^*|\tilde{F}_n)$, where β^* has the same form as $\hat{\beta}$, except that it is based on samples of size n from $\tilde{F}_n$. Generally, the complexity of $\mathcal{G}_n(\beta^*|\tilde{F}_n)$ prevents direct evaluation of its properties, but Monte Carlo resampling from $\tilde{F}_n$ provides a good approximation. Using the joint empirical distribution of the survival times, explanatory variables, and censoring indicators for $\tilde{F}_n$ results in the "nonparametric" resampling plan used by Efron and Tibshirani (1986); i.e., resampling with replacement from the original observations. On the other hand, if a parametric form is assumed for F_β, then

taking $\tilde{F}_n = F_{\hat{\beta}}$ allows "parametric" resampling from the known form of F. In the development below, a semiparametric form of resampling is proposed which uses the incomplete parametric assumption (1), yet allows resampling from a fully-specified distribution.

The proposed method relies on an invariance property of (2) first noted by Kalbfleisch and Prentice (1973). Suppose F_β has conditional survivor function (given $\boldsymbol{x}$),

$$S(t;\boldsymbol{x},\boldsymbol{\beta}) = exp\{-\int_0^t h_0(u)du \cdot g(\boldsymbol{x},\boldsymbol{\beta})\}, \tag{2.1}$$

where the form of g is known and $\boldsymbol{\beta}$ is unknown. Let F'_β be another proportional hazards df with conditional survivor function,

$$S'(t;\boldsymbol{x},\boldsymbol{\beta}) = exp\{-\int_0^t h'_0(u)du \cdot g(\boldsymbol{x},\boldsymbol{\beta})\}. \tag{2.2}$$

Suppose further that both h_0 and h'_0 are nonzero over every open interval in the domain of T. The partial likelihoods (2) are the same for (1) and (2), and the value of $\hat{\boldsymbol{\beta}}$ is not affected by whether we claim the data as originating from F_β or F'_β. In other words, $L(\boldsymbol{\beta})$ remains invariant under monotone increasing transformations of T. This invariance property was previously used by Loughin (1993) in the development of a *residual bootstrap* for $\hat{\boldsymbol{\beta}}$.

A monotone transformation of T of particular interest is

$$Y = F_0(T), \tag{2.3}$$

where F_0 is the distribution function corresponding to an individual with $g(\boldsymbol{x},\boldsymbol{\beta}) = 1$. Then,

$$U \equiv S(T;\boldsymbol{x},\boldsymbol{\beta}) = [1 - F_0(T)]^{g(\boldsymbol{x},\boldsymbol{\beta})}, \tag{2.4}$$

the conditional survivor function of a subject with explanatory variables $\boldsymbol{x}$, has a Uniform(0,1) distribution, and (3) can be written

$$Y = 1 - [U]^{1/g(\boldsymbol{x},\boldsymbol{\beta})}. \tag{2.5}$$

The distribution, survival, and hazard functions of Y are, respectively,

$$\begin{aligned} F_Y(y) &= 1 - (1-y)^{g(\boldsymbol{x},\boldsymbol{\beta})}, \\ S_Y(y) &= (1-y)^{g(\boldsymbol{x},\boldsymbol{\beta})}, \\ h_Y(y) &= g(\boldsymbol{x},\boldsymbol{\beta}) \cdot \frac{1}{1-y}, \end{aligned}$$

$0 < y < 1$. Notice that Y possesses the proportional hazards property, with $h_{0Y}(y) = (1-y)^{-1}$ and relative risk $g(\boldsymbol{x},\boldsymbol{\beta})$. Thus, given a specific form for the relative risk function and a specific value of $\boldsymbol{\beta}$, data can be generated from the proportional hazards distribution F_Y as follows:

1. For $i = 1,\ldots,n$, compute $g(\boldsymbol{x}_i,\boldsymbol{\beta})$;
2. Generate values for n independent Uniform(0,1) variables, $\{u_i\}_{i=1}^n$;
3. Compute $y_i = 1 - u_i^{1/g(\boldsymbol{x}_i,\boldsymbol{\beta})}$, $i = 1,\ldots,n$.

Data generated through these steps are said to be in the *probability scale* of the failure times, and analysis through likelihood (2) can proceed without transforming the data to another time scale.

Assume for the moment that no censoring is present. The invariance of the partial likelihood estimator to monotone transformations of time suggests that the bootstrap can be applied by generating Monte Carlo samples using steps (1)–(3) with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}$ estimated from the original data. In this case, the only stochastic

component in the problem, U, is resampled directly from Uniform(0,1). While this version of the bootstrap resembles a parametric bootstrap in the specification of the proportional hazards property and the estimation of the parameter, it does not require *complete* specification of the form of F_β. Hence, this resampling procedure is called a "semiparametric bootstrap." Note that a parametric bootstrap would further require specifying a particular form for the baseline hazard h_0, in which case a full parametric likelihood could be applied.

In the development of a related resampling plan, the residual bootstrap, of Loughin (1993) resampled estimates of $S(t_i; \boldsymbol{x}_i, \boldsymbol{\beta})$, showing them to be *generalized residuals* by the definition of Cox and Snell (1968). Probability-scale failure times were then created through the transformation (3). The semiparametric bootstrap makes use of the fact that, by the the probability integral transformation, the distribution of $S(T; \boldsymbol{x}, \boldsymbol{\beta})$ is Uniform(0,1), regardless of the underlying baseline hazard function. Thus, "residuals" are resampled parametrically from their true distribution rather than nonparametrically from a set of estimated values.

Extension of this semiparametric bootstrap to a variety of fixed and random censoring schemes is possible. The key is to mimic in the resampling procedure the properties of the censoring mechanisms inherent in the data. Four basic types of censoring mechanisms will be considered here:

1. Censoring based on the ordering of the failures;
2. Censoring distributions depending on $\boldsymbol{x}$;
3. Censoring distributions independent of $\boldsymbol{x}$;
4. Censoring distributions dependent on the *distribution* of T.

Note that the third and fourth mechanisms are both special cases of the second. Each of these mechanisms is assumed to be independent of the failure mechanism. Sensible interpretation of the parameter estimates relies heavily on this assumption. See Kalbfleisch and Prentice (1980), Section 5.2 for a more detailed description of censoring mechanisms.

To introduce censoring into the semiparametric bootstrap, some additional notation is needed. The superscript "^{o}" indicates that an observed time is subject to censoring, so that $T^o = min(T, C)$, where C is the random variable corresponding to time to censoring with distribution function $G_C(c; \boldsymbol{x})$. Using the baseline failure time distribution, the time-to-censoring random variable is transformed to the probability scale of the failure times by

$$W = F_0(C), \tag{2.6}$$

which has distribution function G_W. The probability-scale censoring times are estimated through $\hat{W} = \hat{F}_0(C)$. Although probability-scale failure times can be generated semiparametrically without knowledge of F_0, creation of the corresponding censoring times requires an estimate of F_0 (e.g., using the method of Breslow, 1974).

Corresponding to the definition of U given by (4), define

$$V = (1 - W)^{g(\boldsymbol{x},\boldsymbol{\beta})}, \tag{2.7}$$

and call its d.f. G_V. Recall that U has a Uniform(0,1) distribution, regardless of the baseline distribution, F_0, or the values of the explanatory variables. For this reason, observations of U and V are said to be in the *uniform scale* of the failure times, even though V does not necessarily have a uniform distribution. Values of V are obtained in practice by inserting appropriate estimates for W and $\boldsymbol{\beta}$ into (7).

2.1 Censoring Based on the Ordering of the Failures

Suppose that k of the n subjects in the original sample are observed to fail with ordered failure times $t_{(1)},\ldots,t_{(k)}$, $k<n$. Furthermore, suppose that at m of those times, $t_{(r_1)},\ldots,t_{(r_m)}$, where the values of $r_1,\ldots,r_m$ are fixed in advance, censoring is imposed on the sample. In particular, at time $t_{(r_i)}$, exactly n_i of the subjects still at risk are randomly chosen to be removed from the study. Type II and progressive type II censoring are examples of this method.

Incorporation of this mechanism into the generation of bootstrap samples begins with generating n uncensored probability scale failure times with the semiparametric bootstrap. The times are ordered and the r_1 smallest times are recorded as $y^*_{(1)},\ldots,y^*_{(r_1)}$. Next, n_1 indices are randomly selected from $\mathcal{R}_{r_1+1}$ and their times are censored at $y^*_{(r_1)}$. The failure times for the remaining members of the risk set are ordered, the r_2-r_1 smallest of these times are recorded as $y^*_{(r_1+1)},\ldots,y^*_{(r_2)}$, and n_2 indices are randomly selected form $\mathcal{R}_{r_2+1}$ to be censored at $y^*_{(r_2)}$. This is continued until the k failure times have been selected.

2.2 Censoring Distributions Dependent on Explanatory Variables

The approach taken for generation of censoring times when censoring is random is similar to that of Karrison (1990). In each case, an estimate of the appropriate censoring distribution is required, from which a censoring time is resampled for each observation.

Consider a censoring mechanism which has the potential to vary for individuals with different explanatory variables. This is the case, for instance, when there

are intolerable side effects associated with certain treatments, causing individuals to withdraw from the study. Then different treatment groups will have different censoring distributions. The fact that an individual is censored cannot provide information regarding the remaining time to failure, however. This assumption of independence between the failure and censoring mechanisms is used in the bootstrap procedure.

The semiparametric bootstrap accounts for such censoring mechanisms in the probability scale through the distribution of W. Since $G_C(c;\boldsymbol{x})$ depends arbitrarily on $\boldsymbol{x}$, $G_W(w;\boldsymbol{x})$ also varies with $\boldsymbol{x}$. Hence separate estimates of $G_W(w;\boldsymbol{x})$ are needed for each $\boldsymbol{x}$. Let $\hat{G}_{\hat{W}}(\hat{w};\boldsymbol{x})$ be some estimate of G_W (e.g., the Kaplan-Meier product limit estimator applied to the estimated probability-scale censoring times; see Kalbfleisch and Prentice, 1980). Then corresponding to each $\boldsymbol{x}_i$, $i = 1,\ldots,n$, a probability-scale failure time y_i^* is generated by the semiparametric bootstrap, and a censoring time w_i^* is drawn independently from $\hat{G}_{\hat{W}}(\hat{w};\boldsymbol{x}_i)$. Upon setting $y_i^{o*} = min(y_i^*, w_i^*)$, $i = 1,\ldots,n$, a bootstrap sample with approximately the correct censoring distribution is obtained.

In many problems, the $\hat{G}_{\hat{W}}(\hat{w};\boldsymbol{x})$ are based on a very small number of censored observations in the groups of individuals with common $\boldsymbol{x}$. While the resulting estimate of the censoring distribution may be rather crude, the impact this has on the resampling procedure is small, since censoring plays such a small role in these cases.

Also, in some experiments there is only one individual associated with each value of $\boldsymbol{x}$. This occurs more commonly when explanatory variables are sampled from some population, but may also occur in designed experiments, for example where a wide range of dosage levels are assigned to individuals. In such experiments either all or none of the individuals with a given value of $\boldsymbol{x}$ are censored. In the former case,

estimating censoring distributions as described above implies that only when the resampled failure time is less than the probability-scale estimate of the censoring time can a failure be observed for that $\boldsymbol{x}$. In the latter situation, no censoring is imposed on any resampled failure time for that $\boldsymbol{x}$. It may be more appropriate in some cases to assume some structure to the dependence of C or W on $\boldsymbol{x}$ and estimate $\hat{G}_{\hat{W}}$ by another method.

2.3 Censoring Distributions Independent of Explanatory Variables

Sometimes it may be reasonable to assume that all observations are subject to the same censoring distribution. In such cases, we can write

$$G_C(c;\boldsymbol{x}) = G_C(c), \quad \forall \boldsymbol{x} \tag{2.8}$$

For example, if all individuals are subject to the same fixed endpoint T_c (as in Type I censoring), then $G_C(c;\boldsymbol{x})$ is degenerate at T_c for all $\boldsymbol{x}$. Also, individuals who are lost to follow-up can often be thought of as having censoring distributions independent of $\boldsymbol{x}$. Censoring is then imposed on semiparametric bootstrap samples in the probability scale through a single estimate $\hat{G}_{\hat{W}}$. As in the previous case, failure times $y_1^*, \ldots, y_n^*$ are obtained from a semiparametric bootstrap, and censoring times w_i^*, $i = 1, \ldots, n$ are drawn independently from $\hat{G}_{\hat{W}}$ for all subjects, so that $y_i^{o*} = min(y_i^*, w_i^*)$, $i = 1, \ldots, n$ represents a bootstrap replicate in which failure times are prone to censoring which is independent of explanatory variables. It is straightforward to further modify the resampling scheme of Section 2.2 to situations where censoring depends only on a subset of the variables in $\boldsymbol{x}$.

2.4 Censoring Distributions Dependent on the Distribution of T

An alternative method of estimating censoring distribution can be used when the censoring distribution G is believed to depend on $\boldsymbol{x}$ through the distribution of T. The proportional hazards model for random censoring, due to Koziol and Green (1976) is an example of this dependence. Their model specifies that $1 - G_C(c;\boldsymbol{x}) = [1 - F(t;\boldsymbol{x})]^{\alpha}$, $0 < \alpha < \infty$.

More generally, the dependence of G_C on the distribution of T can be written as $G_C(c;\boldsymbol{x}) = H(1 - F(t;\boldsymbol{x}))$, for some function H. Under the Cox proportional hazards regression model, this is

$$G_C(c;\boldsymbol{x}) = H([1 - F_0(c)]^{g(\boldsymbol{x},\boldsymbol{\beta})}). \tag{2.9}$$

Applying transformations (6) and (7) in (9) yields

$$G_V(v) = H(v),$$

so that the censoring is independent of $\boldsymbol{x}$ in the uniform scale. Hence, an appropriate semiparametric bootstrap proceeds as follows: estimate the uniform-scale censoring times, v_i, $i = 1,\ldots,n$, and their distribution $G_{\hat{V}}$; draw u_i^* $i = 1,\ldots,n$ from Uniform(0,1) and v_i^* $i = 1,\ldots,n$ from $\hat{G}_{\hat{V}}$; take $u_i^{o*} = max(u_i^*, v_i^*)$, $i = 1,\ldots,n$; and apply (5) to the u_i^{o*}'s. Note that the maximum is used here due to the decreasing relationship between Y and U implied by (5).

3. TWO EXAMPLES

To demonstrate the use of the semiparametric bootstrap, the analyses of two familiar data sets with different censoring structures are presented. In both cases, the lone explanatory variable is a fixed constant corresponding to a treatment group assignment. The first example is the childhood leukemia data presented by Cox (1972), consisting of 42 patients randomized into two treatment groups of equal size. The time in remission (in weeks) was subject to random censoring, all of which occurred in the active treatment group, which suggests that the two groups may have completely different censoring mechanisms. Thus, a semiparametric bootstrap with censoring dependent on the explanatory variable is used to estimate the bias and standard error of $\hat{\beta}$ and 90% confidence limits for the parameter β. One thousand bootstrap replicates are collected and analyzed using a loglinear form for the relative risk.

A partial comparison with other bootstrap methods is possible. Efron and Gong (1983) resample independently within each treatment group, which possibly ignores some of the variability in the random assignment of subjects to treatment groups. Efron and Tibshirani (1986) resample triples $\{(t_1^o, x_1, \delta_1)^*, \ldots, (t_n^o, x_n, \delta_n)^*\}$ nonparametrically from the original data $\{(t_i^o, x_i, \delta_i)\}_{i=1}^n$, implicitly taking x to be random. The estimation results and confidence intervals from the application of all of

Table 3.1: Results for the childhood leukemia data

Analysis Method	$\hat{\beta}$	BIAS* [a]	β^*_{BC} [b]	S.E.($\hat{\beta}$)	90% CI Lower	90% CI Upper
Original[c]	1.51	—	—	.41	0.83	2.19
Boot Dep-x[d]	1.58	.07	1.44	.48	0.80	2.25
Boot E&G [e]	na [f]	na	na	.42	0.98	2.35
Boot E&T [g]	na	na	na	.42	1.00	2.39

[a]Bootstrap estimate of bias.
[b]Bias-corrected bootstrap estimate.
[c]Partial likelihood estimation of Cox (1972).
[d]Semiparametric bootstrap with censoring dependent on x.
[e]Bootstrap resampling method of Efron and Gong (1983).
[f]The authors do not provide these values.
[g]Bootstrap resampling method of Efron and Tibshirani (1986).

these methods to the childhood leukemia data are compared in Table 3.1.

The original analysis gives the estimate $\hat{\beta} = 1.51$ (using Breslow's likelihood correction for ties) with a standard error of .41. The semiparametric bootstrap detects an upward bias in the estimate of β and yields a bias-corrected estimate of 1.44. As shown by Loughin (1993), bias correction in this problem provides a noticeable improvement over the original estimator. The standard error estimated by the semiparametric bootstrap is larger than those of the other bootstraps and the asymptotic estimate. In this example, 90% confidence intervals constructed using the various methods all provide essentially the same inference on β. Bootstrap confidence intervals from Efron and Gong (1983) and Efron and Tibshirani (1986) have nearly identical endpoints, although one is constructed to account for bias and the other is not. The bias-corrected percentile interval from the semiparametric bootstrap is the

Table 3.2: Results for the vaginal cancer data

Analysis Method	$\hat{\beta}$	BIAS* [a]	β^*_{BC} [b]	S.E.($\hat{\beta}$)	90% CI Lower	90% CI Upper
Original [c]	-.60	—	—	.35	-1.17	-0.02
Boot Indep [d]	-.62	-.02	-.57	.37	-1.22	0.02
Boot Dep-x [e]	-.61	-.02	-.58	.36	-1.22	0.00
Boot Dep-F [f]	-.62	-.03	-.57	.38	-1.24	-0.03
Boot Resid [g]	-.63	-.04	-.56	.37	-1.18	0.01

[a] Bootstrap estimate of bias.
[b] Bias-corrected bootstrap estimate.
[c] Partial likelihood estimation of Cox (1972).
[d] Semiparametric bootstrap with censoring independent of x.
[e] Semiparametric bootstrap with censoring dependent on x.
[f] Semiparametric bootstrap with censoring dependent on failure time distribution.
[g] Residual Bootstrap of Loughin (1993).

widest of the four, as would be expected from its larger estimate of the variability of $\hat{\beta}$.

The second example is an analysis of the carcinogenesis data in Kalbfleisch and Prentice (1980). Forty rats were randomly assigned to two treatment groups then exposed to a carcinogen. Time (in days) until death due to vaginal cancer was measured for each animal. Two observations in each group were censored at random. No pattern is evident for the censoring, so that any of the three censoring mechanisms discussed in Section 2 may be appropriate. Hence, the analysis is repeated for each mechanism. Also presented is the residual bootstrap analysis in Loughin (1993). In each case, 1000 bootstrap replicates are generated, with $g(x, \beta) = e^{x\beta}$. Results from these methods appear in Table 3.2.

All four bootstrap methods suggest that the original estimate of -.60 is biased

away from zero. Bias-corrected estimates range from -.56 to -.58. The smallest magnitude among these belongs to the estimate from the residual bootstrap, which was found by Loughin to be biased slightly toward zero. Each bootstrap provides a standard error that is slightly larger than the asymptotic estimate of .35. As before, the 90% bias-corrected percentile confidence interval from each bootstrap procedure is wider than the confidence interval based on the usual asymptotic normal approximation. Inverting these intervals for a test of $\beta = 0$ results in differing conclusions, since some intervals include 0 and some do not. However, in all cases 0 is very near the upper limit, so sensible interpretation suggests a borderline significance of the parameter at the .10 level for all methods.

These simple examples require the estimation of only one regression parameter. A previous simulation study by Loughin (1993) suggests that the usual estimation procedures appear to be adequate in such cases. However, it will be seen in Section 4 that when the model contains more than one parameter the semiparametric bootstrap can provide substantial improvement over these procedures in terms of the bias of the parameter estimates, the mean squared error of both the parameter estimates and corresponding variance estimates, and the coverage of the resulting confidence intervals.

4. SIMULATION STUDY

As an objective means of assessing the small-sample precision and accuracy of the proposed methods, results of a Monte Carlo simulation study are presented. In order to allow direct comparison to previous work, the study is modeled to a large degree after the design used in Loughin (1993) for the residual bootstrap. Where the two studies overlap, the same Monte Carlo data sets have been used for the analyses.

Samples of size 50 and 100 are considered. For data without censoring, the uncensored semiparametric bootstrap is compared to the standard method of Cox. Samples are also generated subject to censoring of two different forms. When 20% Type I censoring is imposed, the resampling method with censoring independent of explanatory variables is used. The resampling plan with random censoring depending on the distribution of the failure times is applied to data which are subject to 20% censoring from the model of Koziol and Green. For both censoring types, the more general method is also used which resamples data with different censoring distributions for each different value of $\boldsymbol{x}$.

Two cases are considered. In the first, β has one dimension, corresponding to the effect of a single treatment group indicator x. The samples are split evenly between the two groups, so that x takes on the values 0 and 1 equally often. The parameter value is set at $\beta = 1.5$, as motivated by the childhood leukemia data of

Section 3. The second case represents estimation of a four-dimensional parameter β. The explanatory variable x is a set of four indicators corresponding to membership in one of five treatment groups of equal size. The parameters are assigned the values .2, .4, .6, and .8.

For both cases estimates of the bias and variance of the regression parameter estimator $\hat{\beta}$ maximizing (2) are computed using each of the methods as described above. Bias-corrected estimates of β are used for all bootstraps, as recommended by Loughin (1993). Variance estimates are compared to the unbiased variance estimate provided by the finite sample variance of the 1000 Monte Carlo estimates of $\hat{\beta}$. Mean squared errors (MSEs) are computed for both the parameter and variance estimates. In addition, 90% confidence intervals are constructed for each method of analysis. The standard normal approximation is used to find confidence limits for the non-bootstrap estimates. Bias-corrected percentile limits (Efron, 1981) are computed for all bootstrap methods. In the four-parameter problem, all tabulated values represent averages over the four parameters, for which results were sufficiently similar to allow this summarization.

The occurrence of monotone likelihood sometimes presents a problem to the Monte Carlo examination of the properties of $\hat{\beta}$. As described by Bryson and Johnson (1981), the likelihood (2) is monotone in β whenever certain functions of the explanatory variables are monotone with respect to the ordered failure times. As recommended by these authors, samples in which this occurs are deleted and replaced with new samples in the Monte Carlo trials. The same approach is used when monotone likelihood is detected in a bootstrap replicate. Loughin (1993) warns that heavy incidence of monotone likelihood is an indication that (1) may not be appro-

priate, and that this replacement strategy can yield misleading simulation results. In fact, occurrence of monotone likelihood in more than a few semiparametric bootstrap samples is an indication that the observed data provide little support for a proportional hazards assumption. In this study, sample sizes and censoring amounts are chosen, based on Loughin (1993), to avoid the monotone likelihood problem and allow reliable bootstrap estimation in all cases considered. Even when $n = 50$ and some censoring is present, most of the Monte Carlo trials yield data sets for which the bootstraps experience no monotone likelihood. In the vast majority of trials, the rate of incidence is less than 5/1000, which suggests that very little sampling bias is present.

In all cases, the loglinear relative risk $g(\boldsymbol{x}, \boldsymbol{\beta}) = e^{\boldsymbol{x}\boldsymbol{\beta}}$ is used. For each bootstrap method, 1000 replicates are drawn from each of the 1000 Monte Carlo samples. All computations for both the simulations and the examples are performed in double precision using FORTRAN programs run on DEC workstations at Iowa State University. Maximization of the likelihood is performed initially by a modified Newton-Raphson algorithm, switching to the Powell algorithm if the Newton-Raphson algorithm fails to converge. Pseudo-random uniform numbers are provided by the algorithm of Wichman and Hill (1982).

Table 4.1 contains the average biases and the MSEs for the estimation of β in the one-parameter problem. The bias-corrected bootstrap estimators generally exhibit smaller bias than the standard Cox estimator, except when Type I censoring is present, in which case the biases are all relatively small. Bootstrap methods reduce parameter estimate MSEs for all censoring types by about 9% for $n = 50$ and about 5% for $n = 100$. Where two different resampling plans are applied to the same

data, their outcomes are nearly identical. The same is true of their estimation of $var(\hat{\beta})$, which appears in Table 4.2. Bootstrap methods are slightly conservative for $n = 50$, providing estimates of variance that tend to be about 10% higher than the corresponding finite sample variances. The asymptotic variance estimates tend to be a little low for this sample size. The MSEs for the asymptotic estimates are somewhat better than those for the bootstrap estimates. At $n = 100$ all methods are comparable. Deviations from the finite sample variances are quite small in most cases, and the MSEs for the various methods are generally more similar than in the smaller sample size. There is little difference in the 90% confidence intervals formed by the various methods, as shown in Table 4.3. The standard normal intervals tend to be a little shorter than the bootstrap bias-corrected percentile intervals, but they exhibit no loss in coverage.

Average biases and MSEs for estimating β in the four-parameter problem are provided in Table 4.4. In each case, the bias and MSE of the bootstrap estimators is noticeably smaller than those of the standard estimator. Reductions in MSE due to bootstrap estimation are roughly 15–20% for $n = 50$ and around 10% for $n = 100$. On the average, different resampling methods again provide nearly identical results when applied to the same Monte Carlo data sets.

However, this is not always the case for variance estimation, as seen in Table 4.5. Results vary, depending on the form of censoring involved. For uncensored data, the semiparametric bootstrap provides estimates of variance that are closer to the finite sample variances and have smaller MSEs than the standard asymptotic method for both sample sizes. The method which assumes that the censoring distribution is a function of the time distribution exhibits smaller MSE for variance estimation than

Table 4.1: Simulation results: Biases and mean squared errors for estimation of β in the one-parameter problem

Censoring[a]	Estimation Method	$n = 50$		$n = 100$	
		Bias	MSE	Bias	MSE
None	COX	.031	.138	.039	.071
	BOOT[b]	-.020	.127	.016	.067
Type I	COX	.023	.143	.008	.067
	BOOT INDEP [c]	-.023	.130	-.012	.064
	BOOT DEP-x [d]	-.024	.130	-.012	.064
Random	COX	.043	.175	.037	.084
	BOOT DEP-F [e]	-.013	.158	.011	.079
	BOOT DEP-x	-.014	.157	.011	.079

[a]Censoring fraction is 20%, where applicable.
[b]Bias-corrected semiparametric bootstrap.
[c]Semiparametric bootstrap with censoring independent of x.
[d]Semiparametric bootstrap with censoring dependent on x.
[e]Semiparametric bootstrap with censoring dependent of failure time distribution.

does its more general counterpart when the data are subject to random censoring from the Koziol-Green model. The method for more general censoring mechanisms provides estimates of $var(\hat{\beta})$ that are slightly closer to the finite sample variances than the asymptotic estimates but have slightly larger MSEs. The resampling plan that takes advantage of the particular structure of the censoring distribution does somewhat better on both counts than either of the other two methods. For data with Type I censoring, both the resampling method which assumes censoring is independent of the explanatory variables and the more general method of resampling offer a

slight improvement in MSE over the asymptotic estimator when $n = 50$. However, when $n = 100$, neither bootstrap has MSE as low as that of the asymptotic variance.

In general, bootstrap-based variance estimates tend to be larger than the values they are estimating, while the corresponding asymptotic variance estimates are too small. The 90% confidence interval results, given in Table 4.6 reflect this. The widths of the bootstrap intervals are about 5–10% wider than their normal-based counterparts when $n = 50$. In the larger sample size, the difference is reduced to about 2–6%. In each case, the smallest difference occurs in the Type I censored data. Coverage percentages are generally good for the normal-based intervals when data are subject to Type I censoring, but they are slightly low otherwise. All bootstrap methods exhibit supranominal coverage.

Table 4.2: Simulation results: Sampling variances with mean squared errors in the one-parameter problem

Censoring[a]	Estimation Method	$n = 50$		$n = 100$	
		$\hat{\text{Var}}(\hat{\beta})$	MSE[b]	$\hat{\text{Var}}(\hat{\beta})$	MSE
None	FSV[c]	.138	—	.070	—
	COX	.133	115	.063	13
	BOOT[d]	.153	174	.068	11
Type I	FSV	.143	—	.067	—
	COX	.136	78	.065	6
	BOOT INDEP [e]	.156	155	.069	9
	BOOT DEP-x [f]	.157	159	.069	8
Random	FSV	.173	—	.083	—
	COX	.168	188	.080	19
	BOOT DEP-F [g]	.196	325	.085	25
	BOOT DEP-x	.196	316	.085	24

[a] Censoring fraction is 20%, where applicable.

[b] All MSEs are x10^{-5}.

[c] Variance of the 1000 Monte Carlo estimates of β.

[d] Bias-corrected semiparametric bootstrap.

[e] Semiparametric bootstrap with censoring independent of x.

[f] Semiparametric bootstrap with censoring dependent on x.

[g] Semiparametric bootstrap with censoring dependent of failure time distribution.

Table 4.3: Simulation results: 90% confidence interval widths and coverages in the one-parameter problem

Censoring[a]	Estimation Method	$n = 50$		$n = 100$	
		Width	% Coverage	Width	% Coverage
None	COX	1.19	91.1	0.83	88.4
	BOOT[b]	1.24	90.8	0.84	88.1
Type I	COX	1.21	90.6	0.84	90.2
	BOOT INDEP [c]	1.26	90.0	0.85	90.0
	BOOT DEP-x [d]	1.26	90.4	0.85	90.1
Random	COX	1.34	90.3	0.93	90.7
	BOOT DEP-F [e]	1.40	89.8	0.94	89.9
	BOOT DEP-x	1.40	89.8	0.95	90.5

[a]Censoring fraction is 20%, where applicable.

[b]Bias-corrected semiparametric bootstrap.

[c]Semiparametric bootstrap with censoring independent of x.

[d]Semiparametric bootstrap with censoring dependent on x.

[e]Semiparametric bootstrap with censoring dependent of failure time distribution.

Table 4.4: Simulation results: Biases and mean squared errors for estimation of β in the four-parameter problem

Censoring[a]	Estimation Method	$n = 50$		$n = 100$	
		Bias	MSE	Bias	MSE
None	COX	.043	.277	.035	.126
	BOOT[b]	-.009	.224	.009	.113
Type I	COX	.041	.308	.015	.141
	BOOT INDEP [c]	-.002	.261	-.003	.131
	BOOT DEP-x [d]	-.002	.260	-.003	.131
Random	COX	.066	.347	.018	.149
	BOOT DEP-F [e]	.008	.274	-.008	.134
	BOOT DEP-x	.008	.279	-.010	.135

[a]Censoring fraction is 20%, where applicable.

[b]Bias-corrected semiparametric bootstrap.

[c]Semiparametric bootstrap with censoring independent of x.

[d]Semiparametric bootstrap with censoring dependent on x.

[e]Semiparametric bootstrap with censoring dependent of failure time distribution.

Table 4.5: Simulation results: Sampling variances with mean squared errors in the four-parameter problem

Censoring[a]	Estimation Method	$n = 50$		$n = 100$	
		$\hat{\text{Var}}(\hat{\beta})$	MSE[b]	$\hat{\text{Var}}(\hat{\beta})$	MSE
None	FSV[c]	.275	—	.125	—
	COX	.228	285	.109	30
	BOOT[d]	.289	117	.123	7
Type I	FSV	.306	—	.141	—
	COX	.287	371	.139	25
	BOOT INDEP [e]	.332	335	.148	35
	BOOT DEP-x [f]	.333	339	.148	35
Random	FSV	.342	—	.148	—
	COX	.294	519	.137	32
	BOOT DEP-F [g]	.369	373	.153	22
	BOOT DEP-x	.376	571	.154	35

[a] Censoring fraction is 20%, where applicable.

[b] All MSEs are x10^{-5}.

[c] Variance of the 1000 Monte Carlo estimates of β.

[d] Bias-corrected semiparametric bootstrap.

[e] Semiparametric bootstrap with censoring independent of x.

[f] Semiparametric bootstrap with censoring dependent on x.

[g] Semiparametric bootstrap with censoring dependent of failure time distribution.

Table 4.6: Simulation results: 90% confidence interval widths and coverages in the four-parameter problem

Censoring[a]	Estimation Method	$n = 50$ Width	$n = 50$ % Coverage	$n = 100$ Width	$n = 100$ % Coverage
None	COX	1.57	87.8	1.09	88.2
	BOOT[b]	1.74	91.9	1.15	90.0
Type I	COX	1.76	90.0	1.23	90.5
	BOOT INDEP [c]	1.87	90.9	1.26	91.1
	BOOT DEP-x [d]	1.87	91.3	1.26	91.2
Random	COX	1.78	88.4	1.22	89.2
	BOOT DEP-F [e]	1.96	91.6	1.28	90.6
	BOOT DEP-x	1.98	92.0	1.28	91.1

[a] Censoring fraction is 20%, where applicable.

[b] Bias-corrected semiparametric bootstrap.

[c] Semiparametric bootstrap with censoring independent of x.

[d] Semiparametric bootstrap with censoring dependent on x.

[e] Semiparametric bootstrap with censoring dependent of failure time distribution.

5. DISCUSSION

The semiparametric bootstrap and the residual bootstrap of Loughin (1993) are closely related procedures for estimation of regression parameters in proportional hazards models. The censoring distribution for the residual bootstrap is assumed to follow (9). Both resampling methods appear to provide an improvement over standard estimation procedures for this problem when when this assumption is met. A direct comparison to the corresponding semiparametric resampling plan shows that the semiparametric bootstrap creates estimates of β that are less biased but have higher MSE, estimates of $var(\hat{\beta})$ that are slightly higher but have smaller MSE, and confidence intervals that are slightly wider but have similar coverage.

Part of the reason that the residual bootstrap estimator exhibits more bias may be that it produces tied failure times in the bootstrap samples. Since the semiparametric bootstrap generates continuous data, it does not induce ties in resampled data even when ties are present in the original data. When the original data contains some tied failure times, some adjustment to the semiparametric bootstrap may be needed to account for the bias they may cause in $\hat{\beta}$.

As mentioned in the discussion of the residual bootstrap in Loughin (1993), the matter of the best bootstrap confidence interval for this problem remains open. It is apparent that the bias-corrected percentile limits of Efron (1981) are too wide.

Plots not presented here reveal that estimates of $var(\hat{\beta})$ increase as $\hat{\beta}$ moves away from zero. Percentile limits are not intended to adapt to such situations. The accelerated bias-corrected intervals of Efron (1987) implicitly provide some variance stabilization, and hence might perform better for this problem. Also, bias-correction of parameter estimates is known to be an effective means of reducing both the bias and the variability of estimates of β. Unpublished simulation results show that the *ad hoc* method of using bootstrap bias-corrected estimates of β with the asymptotic estimate of variance provides coverage that meets or exceeds the nominal level while maintaining the shorter widths that the standard normal intervals enjoy.

As the simulations of Section 4 indicate, for both sample sizes, bias and MSE of the Cox estimator are lower for data subject to 20% Type I censoring than for 20% randomly-censored data. Reasons for this are not entirely clear. A more thorough investigation into the properties of this estimator under a wider variety of censoring types may provide a better understand sources of its bias, enabling the development of resampling plans to address this..

Andersen and Gill (1982), and Bailey (1983) have established the consistency of the estimators of β and $var(\hat{\beta})$ resulting from the maximization of the partial likelihood when explanatory variables are not random. When there is no censoring, proof of consistency of the corresponding semiparametric bootstrap estimators follows immediately from their work, since the resampling procedure is equivalent to parametric resampling from $F_{\hat{\beta}}$. However, consistency is not so easily established when censoring is present. Large-sample theory for the methods of Efron and Gong (1983), Efron and Tibshirani (1986), or Loughin (1993) must deal with the presence of tied failure times in the bootstrap samples. The existing large-sample theory for

the proportional hazards model assumes sampling from a continuous distribution, where ties cannot occur.

When data are censored, the assumption of independent censoring and failure mechanisms is crucial to the generation of the semiparametric bootstrap samples. In some problems, however, the independent censoring assumption may not be reasonable. Instead censoring might be viewed as another form of failure, so that the nonparametric resampling methods of Efron and Gong (1983), Efron and Tibshirani (1986), and Loughin (1993), may be more appropriate, since they resample censoring indicators along with failure times or residuals. These methods are equivalent to resampling procedures for the multivariate failure problem proposed by Loughin (1993).

BIBLIOGRAPHY

Altman, D. G. and Andersen, P. K. (1989). Bootstrap Investigation of the Stability of the Cox Regression Model. *Statistics in Medicine*, **8**, 771–783.

Andersen, P. K. and Gill, R.D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Approach. *Annals of Statistics*, **10**, 1100–1120.

Bailey, K. R. (1983). The Asymptotic Joint Distribution of Regression and Survival Parameter Estimates in the Cox Model. *Annals of Statistics*, **11**, 39–48.

Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, **30**, 89–99.

Bryson, M. C. and Johnson, M. E. (1981). The Incidence of Monotone Likelihood in the Cox Model. *Technometrics*, **23**, 381–383.

Chen, C. H. and George, S. L. (1985). The Bootstrap and Identification of Prognostic Factors via Cox's Proportional Hazards Regression Model.

Statistics in Medicine, **4**, 39–46.

Cox, D. R. (1972). Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–202.

Cox, D. R. (1975). Partial Likelihood. *Biometrika*, **62**, 269–276.

Cox, D. R. and Snell, E. J. (1968). A General Definition of Residuals (with discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 248–265.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, **7**, 1–26.

Efron, B. (1981). Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics*, **9**, 139–172.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, **82**, 171–185.

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, Jackknife, and Crossvalidation. *American Statistician*, **37**, 36–48.

Efron, B. and Tibshirani, R. J. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical*

Science, **1**, 54–77.

Farewell, V. T. and Prentice, R. L. (1980). The Approximation of Partial Likelihood with Emphasis on Case Control Studies. *Biometrika*, **67**, 273–278.

Freedman, D. A. (1981). Bootstrap Regression Models. *Annals of Statistics*, **9**, 1218–1228.

Johnson, M. E., Tolley, H. D., Bryson, M. C., and Goldman, A. S. (1982). Covariate Analysis of Survival Data: A Small Sample Study of Cox's Model. *Biometrics*, **38** 685–698.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal Likelihoods Based of Cox's Regression and Life Model. *Biometrika*, **60**, 267–278.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Karrison, T. (1990). Bootstrapping Censored Data with Covariates. *Journal of Statistical Computation and Simulation*, **36**, 195–207.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises Statistic for Randomly Censored Data. *Biometrika*, **63**, 465–474.

Loughin, T. M. (1993). Bootstrap Applications in Proportional Hazards Models. *Ph.D. Dissertation, Iowa State University*

Wichman, B. A. and Hill, I. D. (1982). An efficient and portable pseudo-random number generator. *Applied Statistics*, **31**, 188–190.

Wu, C. F. J. (1986). Jackknife, Bootstrap, and Other Resampling Methods in Regression. *Annals of Statistics*, **14**, 1261–1295.

GENERAL SUMMARY

The papers in this thesis describe techniques for handling the very important statistical problem of the analysis of event times. When a single event time is observed for each study subject, Cox's proportional hazards model often provides a good description of the relationship between the event time and a given set of explanatory variables. The model is flexible, both in admitting numerous forms for this relationship and in allowing a wide variety of structures for the distribution of events in time. The two examples given in Papers I and III are typical of survival studies in medical applications, where the proportional hazards model is widely applied. The improvements made in these papers may thus impact analyses of research results wherever the model is used.

When multiple events are observed on each subject, there is little consensus on how to model the relationship between the times and the explanatory variables. The complicating factor of association among event times makes testing the applicability of many of the available multivariate models difficult. In many studies this association is largely a nuisance to the estimation of the effects of the explanatory variables. Even where there is interest in making comparisons of those effects corresponding to different event types, the practitioner often has little concern for the mathematical structure of the association. The Independence Working Model approach used by

several authors and applied in Paper II avoids the specification of a structure for this association. Instead, models for each event type are chosen based on their applicability for that event type alone, without consideration of models needed for other event types. Hence, the method proposed in Paper II for estimation of the relationships between explanatory variables and different event times has wide applicability.

The resampling procedures developed in this thesis represent a change in the approach to bootstrapping in a variety of problems. In the first and third papers, the invariance of the partial likelihood of Cox (1972) to monotone increasing transformations allows generation of bootstrap replicates in a different scale from that of the original data. This special property of the estimation procedure may be present in other estimation problems, and in these cases similar resampling procedures might be available. The second paper addresses a problem in which the model proposed for the data is incomplete in that the margins of the distribution for the data are be specified completely (up to a set of unknown parameters), while the association among the margins is left to vary freely. Partial parametric specifications are common in problems in which the primary interest is in assessing the effects of some explanatory variables on one or multiple responses. The work by Simonoff and Tsai (1988) on jackknifing and bootstrapping quasilikelihood estimators is another example of an application of resampling methods to problems of this type. There is room for a great deal more work in this area.

In the discussions at the end of each paper, suggestions are given for further research directions. The most immediate need may be identifying the best procedures for conducting bootstrap-based inference on β, since new methods in statistics cannot gain favor in application if the resulting inferential procedures are not well de-

veloped. The confidence interval section of the discussion of Paper III suggested that some use might be made of the bias-corrected estimates β^*_{BC}. Such estimates from the resampling plans proposed in all three papers provide an improvement over the standard method of estimating β, as measured by the mean squared error. However, no estimate of the standard error of β^*_{BC} is immediately available.

This problem could possibly be handled by an *iterated bootstrap* (see, e.g., Hall, 1992), where the bootstrap estimator is itself bootstrapped. In other words, from each bootstrap sample taken from the original data, new sets of observations are resampled, and the bootstrap estimator is computed on each of these data sets. Of course, this further increases the computational intensiveness, literally by powers of two. Also, except in the case of a semiparametric bootstrap, an increase in the proportions of tied failure times will occur. Perhaps the greatest drawback, however, is the possibility that monotone likelihood will make estimation in the iterated bootstrap unreliable. To better understand this problem, recall the one-parameter simulations with $n = 24$ and no censoring from Paper I. Only 6 of the 1000 Monte Carlo trials are replaced due to monotone likelihood. Using the detect-and-replace strategy of Bryson and Johnson, this is not likely to have a profound effect on the resulting Monte Carlo estimates. However, the ensuing bootstraps for many of the Monte Carlo Trials are plagued with monotone likelihood problems. This situation could occur in an analogous manner in the iterated bootstrap. However, in most cases where large proportions of replicates may produce monotone likelihoods, the parameter values are sufficiently large that any reasonable inferential procedure is likely to be adequate, at least for testing purposes.

Another issue requiring further development is the handling of experiments in

which both fixed and random explanatory variables are used. It is often the case in practice that some concomitant information, such as age of a patient, is available and could play an important role in the patient's survival experience. Such variables are not considered fixed by study design, unless some sort of stratified sampling is used. A bootstrap resampling plan might be developed under the assumption that the explanatory variables and the residuals follow some joint distribution on an appropriate support set.

The estimator used for the residuals in the residual bootstrap is brought into question in Paper I. As discussed in the General Introduction, several different estimators of the conditional survivor function might provide more nearly uniformly-distributed variates for resampling. This may reduce the bias observed in the bias-corrected residual bootstrap estimates of β. Another approach is suggested by the work of Cox and Snell. They derived adjustments to their generalized residuals, based on second-order expansions of the maximum likelihood estimators of the residuals, to provide estimates with correct bias and variance to order n^{-1}. Application of their corrections to the conditional survivor function estimates may be useful.

BIBLIOGRAPHY

Altman, D. G. and Andersen, P. K. (1989). Bootstrap Investigation of the Stability of the Cox Regression Model. *Statistics in Medicine*, **8**, 771–783.

Andersen, P. K. and Gill, R.D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Approach. *Annals of Statistics*, **10**, 1100–1120.

Bailey, K. R. (1983). The Asymptotic Joint Distribution of Regression and Survival Parameter Estimates in the Cox Model. *Annals of Statistics*, **11**, 39–48.

Barlow, W. E. and Sun, W. H. (1989). Bootstrapped Confidence Intervals for the Cox Model using a Linear Relative Risk Form. *Statistics in Medicine* **8**, 927–935.

Beran, R. (1982). Estimating Sampling Distributions: The Bootstrap and Competitors. *Annals of Statistics*, **10**, 212–225.

Bickel, P. J. and Freedman, D. A. (1981). Some Asymptotic Theory for the

Bootstrap. *Annals of Statistics*, 9, 1196–1217.

Breslow, N. (1972). Comment on "Regression Models and Life Tables." *Journal of the Royal Statistical Society, Series B*, **34**, 216–217.

Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, **30**, 89–99.

Bryson, M. C. and Johnson, M. E. (1981). The Incidence of Monotone Likelihood in the Cox Model. *Technometrics*, **23**, 381–383.

Chen, C. H. and George, S. L. (1985). The Bootstrap and Identification of Prognostic Factors via Cox's Proportional Hazards Regression Model. *Statistics in Medicine*, **4**, 39–46.

Clayton, D. G. (1978). A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*, **65**, 141–151.

Clayton, D. and Cuzick, J. (1985). Multivariate Generalizations of the Proportional Hazards Model (with discussion). *Journal of the Royal Statistical Society, Series A*, **148**, 82–117

Costanza, M. C. and Nichola, P. S. (1982). Effect of Random Censoring on

Cox-Breslow Survival Methods: A Simulation Study. *ASA Proceedings from the Section on Statistical Computation*, 258–262.

Cox, D. R. (1972). Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–202.

Cox, D. R. (1975). Partial Likelihood. *Biometrika*, **62**, 269–276.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* New York: Chapman and Hall.

Cox, D. R. and Snell, E. J. (1968). A General Definition of Residuals (with discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 248–265.

Crowder, M. (1989). A Multivariate Distribution with Weibull Connections. *Journal of the Royal Statistical Society, Series B*, **58**, 93–107.

Crowley, J. and Hu, M. (1977). Covariance Analysis of Heart Transplant Survival Data. *Journal of the American Statistical Association*, **77**, 27–36.

Crowley, J. and Storer, B. E. (1983). Comment on "A Reanalysis of the Stanford Heart Transplant Data." *Journal of the American Statistical Association*, **78**, 277–281.

DeGroot, M. H. (1975). *Probability and Statistics.* Reading, Massachusetts: Addison-Wesley.

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, **72**, 557–565.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, **7**, 1–26.

Efron, B. (1981a). Censored Data and the Bootstrap. *Journal of the American Statistical Association*, **76**, 312–319.

Efron, B. (1981b). Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics*, **9**, 139–172.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. *SIAM monograph*, **38**, CBMS-NSF.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, **82**, 171–185.

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, Jackknife, and Crossvalidation. *American Statistician*, **37**, 36–48.

Efron, B. and Tibshirani, R. J. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, **1**, 54–77.

Farewell, V. T. and Prentice, R. L. (1980). The Approximation of Partial Likelihood with Emphasis on Case Control Studies. *Biometrika*, **67**, 273–278.

Feigl, P. and Zelen, M. (1965). Estimation of Exponential Survival Probabilities with Concomitant Information. *Biometrics*, **21**, 826-838.

Freedman, D. A. (1981). Bootstrap Regression Models. *Annals of Statistics*, **9**, 1218–1228.

Freund, J. E. (1961). A Bivariate Extension of the Exponential Distribution. *Journal of the American Statistical Association*, **56**, 971–977.

Gumbel, E. J. (1960). Bivariate Exponential Distributions. *Journal of the American Statistical Association*, **55**, 698–707.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer-Verlag.

Hougaard, P. (1987). Modeling multivariate Survival. *Scandinavian Journal of Statistics*, **14**, 291–304.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimated under Nonstandard Conditions. *Proceedings of the Fifth Berkeley Symposium*, 221-233.

Huster, W. J., Brookmeyer, R., and Self, S. G., (1989). Modeling Paired Survival Data with Covariates. *Biometrics*, **45**, 145–156.

Johnson, M. E., Tolley, H. D., Bryson, M. C., and Goldman, A. S. (1982). Covariate Analysis of Survival Data: A Small Sample Study of Cox's Model. *Biometrics*, **38** 685–698.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal Likelihoods Based of Cox's Regression and Life Model. *Biometrika*, **60**, 267–278.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of Likelihood Methods to Models involving a Large Number of Parameters (with discussion). *Journal of the Royal Statistical Society, Series B*, **32**, 175–208.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457–481.

Karrison, T. (1990). Bootstrapping Censored Data with Covariates. *Journal of Statistical Computation and Simulation*, **36**,

Kay, R. (1977). Proportional Hazard Regression Models and the Analysis of Censored Survival Data. *Applied Statistics*, **26**, 227–237.

Klein, J. P., Keiding, N., and Kamby, C. (1989). Semiparametric Marshall-Olkin Models Applied to the Occurrence of Metastases at Multiple Sites after Breast Cancer. *Biometrics*, **45** 1073–1086.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises Statistic for Randomly Censored Data. *Biometrika*, **63**, 465–474.

Lagakos S. W. (1980). The Graphical Evaluation of Explanatory Variables in Proportional Hazards Regression Models. *Biometrika*, **68**, 93–98.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* New York: Wiley.

Lin, D. Y. and Wei, L. J. (1989). The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*, **84**, 1074–1078.

Marshall, A. W. and Olkin, I. (1967). A Multivariate Exponential Distribution.

Journal of the American Statistical Association, **62**, 30–44.

Miller, R. G. (1974). The Jackknife—A Review. *Biometrika*, **61**, 1–15.

Neter, J., Wasserman, W., and Kutner, M. H. (1985). *Applied Linear Statistical Models.* Homewood, Illinois: Richard D. Irwin.

Oakes, D. (1972). Comment on "Regression Models and Life Tables." *Journal of the Royal Statistical Society, Series B*, **34**, 208.

Oakes, D. (1982). A Model for Association in Bivariate Survival Data. *Journal of the Royal Statistical Society, Series B*, **44**, 414-422.

Peto, R. (1972). Comment on "Regression Models and Life Tables. *Journal of the Royal Statistical Society, Series B*, **34**, 205.

Royall, R. M. (1986). Model Robust Confidence Intervals using Maximum Likelihood Estimators. *International Statistical Review*, **54**, 221–226.

Schenker, N. (1985). Qualms about Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, **80**, 360–361.

Simonoff, J. S. and Tsai, C.- L. (1988). Jackknifing and bootstrapping quasi-likelihood estimators. *Journal of Statistical Computation and Simulation*,

30, 213–232.

Singh, K. (1981). On the Asymptotic Accuracy of Efron's Bootstrap. *Annals of Statistics*, **9**, 1187–1195.

Tsiatis, A. A. (1981). A Large Sample Study of Cox's Regression Model. *Annals of Statistics*, **9**, 93–108.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, **84**, 1064–1073.

Wichman, B. A. and Hill, I. D. (1982). An efficient and portable pseudo-random number generator. *Applied Statistics*, **31**, 188–190.

Wu, C. F. J. (1986). Jackknife, Bootstrap, and Other Resampling Methods in Regression. *Annals of Statistics*, **14**, 1261–1295.